IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(RESEARCH ARTICLE)

A Comprehensive Study of Datasets for Research Works in IoT Networks

SUPARNA N * and MANJAIAH D H

Department of PG Studies and Research in Computer Science, Mangalore University, Mangalore, India.

International Journal of Science and Research Archive, 2024, 13(01), 2652–2662

Publication history: Received on 02 September 2024; revised on 13 October 2024; accepted on 15 October 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.1.1941

Abstract

Internet of Things is a trending subject in the field of Computer Science and Information Technology. When a research scholar tries to find a problem statement in the arena of IoT, he or she probably Google the phrase "Security Issues in Internet of Things". On this day August 22, 2024, when we searched the Google with this phrase, we got about 48,40,00,000 results in 0.46 seconds. When we searched Google Scholar with the same phrase, we got about 40,70,000 results (0.52 sec). This shows that lakhs of researchers are toiling to find solutions for Security Issues in IoT. There is a new and trending way of research that is finding solutions for issues in IoT using Machine Learning and Deep Learning Models. ML and DL algorithms need large sets of Data to train and test the algorithms. This made us to take up a study on available Datasets for this research topic. When we made review of literature, in this arena of study, we found that researchers are working on only a few datasets, though numerous datasets are available freely on the internet for academic purposes. Hence in this paper we made a comprehensive study of Datasets available for research works in *Internet of Things.*

Keywords: Datasets; Machine Learning; Kaggle; Papers with code; Deep learning; Figshare.

1. Introduction

According to Cambridge Dictionery, a Dataset is defined as "A collection of separate sets of information that is treated as a single unit by a computer". If we analyse this definition we can understand that a dataset represents a particular category of data which is organized into rows and columns such that a set of information can be generated from a dataset. A dataset is basically a collection of records or a file of records where the characteristics of each record are the same. The charcteristics or the attributes or the fieldnames are called features of the dataset. Datasets generally represent a unique body of work and cover only one topic such as medical record of breast cancer patients, record of vineyards affected by Downy Mildew disease, record of temperatures of a city etc. Datasets can be reffered just by its name without any details of its storage location or any other details and thus datasets can be easily cataloged.

The most common and readily available datasets are log files of any unbroken and uninterrupted input/output sessions, time based analytical type of reports from the web, market analyses, server logs etc. Many people think that creating a dataset is a tedious and daunting task which is a myth. A dataset can be created from any type of use case examples or real-world projects. But before creating a dataset, the researchers must have a clear and well-defined purpose that gives an acceptable standard for the newly created dataset. If we would like to create a dataset of images, it should be in an appropriate format such that Machine Learning or Deep learning algorithms can use it to develop image classifiers. A dataset created from the input-output data of the billing software at a supermarket, can be easily adopted to analyze and train recommendation systems. Datasets can be categorized with respect to different concepts. According to type of the data stored in datasets, they can be classified as image, multivariate, sequential, spaciotemporal, tabular, text and time series etc. With respect to the feature type, datasets are of three types namely Numerical, Categorical and Mixed type.

* Corresponding author: SUPARNA N

2. Review of Literature

2.1. A Review on Cyber Security Datasets for Machine Learning Algorithms [1]:

Ozlem Yavanoglu and Murat Aydos, in this article compared the most commonly used datasets on Cyber Security. In the Literature Review Section, the authors studied several papers which dealt with common Cyber Threats such as DDoS Attack, Botnet Attack, Intrusion Detection, HTTP attacks, and Cloud Security. Among these papers nearly 8 works have taken KDD CUP 99 dataset for their study. The authors of [1] studied and tabulated the features, quantity of data and composition of UNSW-NB15, ADFD-LD, CTU-13, HTTP CSIC 2010, ISOT, ECML-PKDD 2007 and KDD CUP 99 datasets.

2.2. Dataset Search: A Survey [2]

Adriane Chapman et al have studied the different ways of searching for datasets on the internet. The authors surveyed the state-of-the-art research and commercial systems and presented the challenges and open questions about dataset search. They state that there is a disconnect between what datasets are available, what dataset a user needs, and what datasets a user can find, trust and is able to use. There are two categories of dataset search namely Basic dataset search and constructive dataset search. In this survey, we can find appropriate examples to understand both categories. We also get an abstract view of the search process, comprising of querying, query processing, data handling and results presentation, alongside approaches to each step subcategories like Tabular search, Entity-centric search, information retrieval etc. The dataset Basic Search implementations are mainly of three types Centralized, Decentralized and Constructive Search. As a concluding remark, the authors emphasized that there is a mounting demand for specialized search engines for datasets of all sorts along with appropriate, easy to use tools[2].

2.3. New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Dataset[3]

Nour Moustafa et al at the Cyber Range Labs of University of New South Wales, Australia created a testbed to generate a new dataset called TON-IoT Dataset. Abdullah Alsaedi et al published another paper [4] on the same dataset and highlighted the relevance IIoT networks in the dataset. This dataset incorporates both normal sensor measurement data as well as various types of attacks targeting IIoT applications[5].

2.4. Survey of Intrusion Detection Systems: Techniques, Datasets And Challenges[6]

Ansam Khraisat et al in their survey of IDS, listed the features and limitations of existing datasets such as DARPA / KDD Cup99, CAIDA, NSL-KDD, CICIDS 2017, ISCX 2012, ADFA-LD and ADFA-WD. A Comparison table of results achieved by various methods on publicly available IDS datasets was presented in the study[6].

2.5. A Comprehensive Survey of Video Datasets for Background Subtraction [7]

In this comprehensive study the authors Rudrika Kalsotra and Sakshi Arora listed the details of video datasets dedicated to background subtraction for detection of moving objects.

2.6. Datasets for Large Language Models: A Comprehensive Survey[8]

Yang Liu et al collated all the datasets available for Large Language Models to gain insights into their status and future trends, consolidated and categorized the fundamental aspects of LLM datasets from five perspectives viz., Pre-training Corpora, Instruction Fine-tuning Datasets, Preference Datasets, Evaluation Datasets, Traditional Natural Language Processing (NLP) Datasets. This is a massive study with a 180 pages report has been developed with a precise architecture.

2.7. Dataset Search Platforms -An Overview

Most of the Supervised Machine learning algorithms are showing high levels of accuracy, when trained and tested upon large amounts of labelled data. In recent years research works using ML and DL Algorithms on several types of datasets are increasing exponentially. The various types and formats of datasets are available on the internet from various arena like marketing, shares trading, banking, medical, sports, environment, smart city projects, scientific, IoT projects, social media etc. Therefore novel, innovative and highly beneficial research works are being hypothesized and datasets are being rummaged from the vast internet with various search phrases. Use of vague search phrases while searching for datasets and ignorance about the quantity, quality, source, purpose and features of datasets make the researchers to struggle at the initial stages of research.

Though predictions and detections of various issues using Machine Learning and Deep Learning algorithms are giving accurate results, this field is facing critical bottlenecks in data collection.

Data collection is a comprehensive process of choosing the dataset, labelling, cleansing, visualizing, analyzing and feature engineering. As previously mentioned, the accuracy of predictions or detections increase only if substantial quantity of labelled dataset is used for training. In widespread applications like cyber-attacks, computer vision, image processing. signal processing etc., massive quantity of data is amassed since decades. Whereas, while developing machine learning models for new applications like VANETs, smart factories, smart cities etc., there is little, or no labelled training data available to start with. For example, in a smart city project, if there is a need for development of detection model to detect garbage throwers using machine learning, there will not be training data to start with. To overcome these challenges, researchers explore alternative datasets such as lab-based simulation techniques to generate training datasets, synthetic datasets and crowdsourced datasets etc. These issues demand a need of accurate and scalable datasets, and hence motivated us to conduct a comprehensive survey of the datasets available for the research works. The article published in IEEE international conference on big data (big data), 2017 by O Yavanoglu and M Aydo [1] which gives an extensive study of datasets available for research works in Cyber Security issues is another motivation for bringing up this study. The above-mentioned paper has gained 188 citations as on 25, August 2024.

In the Table I, we have presented a non-exhaustive list of sources or search platforms for datasets where thousands of datasets are available under various types of licenses.

2.7.1. KAGGLE

Anthony Goldbloom founded Kaggle in 2010 as a data-hub for data science and machine learning geeks. The exponential growth of Data Science, ML, DL, AI and Robotics has made Kaggle as a second to none platform for researchers, students and data analysts. Now owned by Google LLC, this exhaustive dataset repository dispenses millions of datasets free of cost. It covers wide range of topics, fields, subjects, domains and issues.

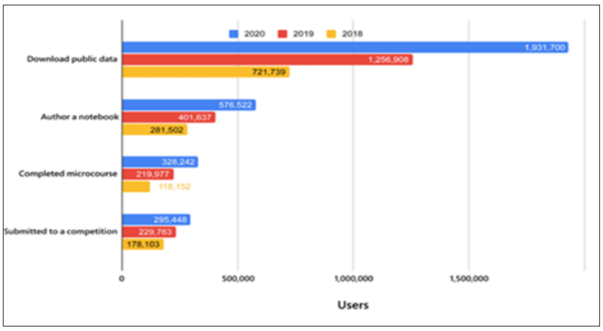

Figure 1 Usage of Kaggle in 2020, 2019, and 2018

Kaggle comprises of seven different aspects namely Competitions, Datasets, Notebooks, Discussion Forums, Kernels, Community and Networking. Konrad Banachewicz and Luca Massaron have authored The Kaggle Book[9] which gives thorough information about the massive repository.

The founder and CEO of Kaggle tweeted a bar graph which showed the statistics about the popularity of Kaggle. The graph in Figure 1, depicted the number of public datasets downloaded, number of users who created public Notebook in Python and the number of users who used all other aspects of Kaggle.

2.7.2. Paperswithcode

This platform which aims at bridging the gap between academic research and practical implementations was founded by Sandeep Subramanian and Siddhartha Kiran. Mainly targeting the researchers and learners in the arena of ML, DL and AI, this community project allows its users to access both research papers and their associated code. Paperswithcode is a state-of-the-art repository which provides benchmarks, tasks, datasets and research articles with codes under the various categories of subjects like Computer Vision, NLP, Medical, Music, Games etc. The platform is well-organized with respect to various ML algorithms and catalogs, under 134 categories. The datasets collection can be searched concerning Best Match, Newest and Most Cited. More than 150 datasets get listed when the search phrase “IoT Security Threats” is used for searching datasets. The researchers can find trendy papers and datasets like TON_IoT, BOT_IoT, CIC-IoT etc.

2.7.3. Quantumstat

Founded by Rickey Costa, QuantumStat provides business solutions for NLP and AI Initiatives, combined from different sources. As stated by the portal itself, this is an NLP Model Forge and dispenses the NLP tasks like Sequence Classification, Token Classification, Question Answering, Summarization, CommonSense, Natural Language Inference, Automatic Text Generation, Conversational, Machine Translation, Text-to-Speech. Every NLP aficionado may be a regular visitor to Quantumstat as it offers thousands of models, datasets, and Colab notebooks. The systematic organization of the portal makes it distinct as it provides a download link for datasets, appropriate pdf article that refers the task and its Google Colab files. Nearly 200 NLP Datasets and 8000 Repos are available in this unique repository.

2.7.4. Google Dataset Search

To foster a data sharing eco system such that the researchers and data publishers will follow the best practices while publishing their datasets and to assist the authors and academicians who use Google Scholar, Natasha Noy, a research Scientist at Google AI along with other AI scientists developed Google Dataset Search and released the beta version in September 2018. After some modifications and enhancements, a full-fledged version was out for public use in January 2020. Being a dedicated search engine for datasets, GDS indexes more than 45 million datasets from nearly 13,000 websites. In their article[10] Natasha Noy et al highlighted that while developing GDS they have faced several challenges to aggregate, normalize and reconcile metadata of datasets mainly because the dataset owners and publishers publish semantically enhanced metadata in their own sites. The metadata of a dataset must have salient features such as its title and description, provider, spatial and temporal coverage, distribution, abstract, author, contributor, country Of Origin, Data Catalog and so on. The developers of GDS have used the metadata specification given in schema.org.

Omar Benjelloun et al in the article [11], have described GDS as a snapshot of datasets on the Web. The authors have analyzed the corpus of dataset metadata used in Google's Dataset Search. Through this analysis work, authors have taken up the challenge of improving metadata quality by automated techniques, feedback to publishers, developing interactive tools to create and validate dataset metadata and crowdsourcing. Syed Ali Hussain studied GDS and published his paper [12] in American Journalism. He states that GDS is user-friendly dataset search portal and understands the user's needs. Its biggest advantage is that it organized the list of datasets that were fragmented onto different sites. Even though the search phrase contains only single word, GDS dispenses 100+ results and displays the features of each dataset such as Explore at, Dataset updated, Dataset provided by, Authors, License, Time period covered, Area covered, Description[12].

2.7.5. OPENML

Founded by Joaquin Vanschoren a Machine learning professor at TU Eindhoven, OpenML is being built by an open source community who believe that ML research should be open, well organized and easily accessible and reusable. Jan N. van Rijn et al in [13] present OpenML, a novel open science platform that provides easy access to machine learning data, software and results to encourage further study in Machine Learning. This features a web API integrated with popular ML tools like KNIME, RapidMiner, Weka and R Packages. This is an open platform which shares datasets, algorithms and experiments with the tagline, 'From the ML Community to the ML Community'. All codes of the OpenML project carries the BSD-3 Clause license.

OpenML has a set of 27 Benchmarking suites for classification, timing attacks, padding attacks, regression, generalization error etc. Bernd Bisch et al developed curated and practical benchmarking suite for classification namely the OpenML Curated Classification benchmarking suite 2018 (OpenML-CC18)[14].

Collaborative and open platforms always need benchmarking suits because they allow the public to share data. OpenML is also an open and collaborative platform where users share new datasets, their own models and experiments. An OpenML benchmarking suite is a set of OpenML tasks carefully picked to evaluate algorithms under a precise set of conditions. It organizes the repository contents based on fundamental building blocks viz., data, task to be solved, task type, target feature, the evaluation procedure, target performance metric, pipeline, and the run that contains experimental results. Bernd Bischl et al designed OpenML benchmarking suites which allow researchers to compile and publish well-defined collections of curated tasks and datasets, and collect benchmarking results from many scientists in a single place[14]. Giuseppe Casalicchio et al in [15] presented an R package to interface with the OpenML platform. This OpenML R package is an interface to interact with the OpenML server directly from within R. Users can retrieve data sets, tasks, flows and runs from the server and create and upload their own[15]. Besim Bilalli et al in their work [16] have proposed an easy to use application for retrieving different meta-data from OpenML.

Table 1 Datasets Search Platforms

Sl. No	Organization Name	Website URL	No. of Datasets Available	License
1	Kaggle	www.kaggle.com	370K	Common Licenses
2	PaperswithCode	https://paperswithcode.com/datasets	10,511(as on 5th, Sep 2024)	CC BY-SA
3	QuantumStat	https://quantumstat.com/ https://index.quantumstat.com/	Nearly 200 NLP Datasets 8,000 Repos	Creative Commons
4	Figshare	https://figshare.com/	NA	NA
5	Google Dataset Search	https://datasetsearch.research.google.com/	45 million Datasets are indexed	Various
6	Stanford University	https://snap.stanford.edu/data/	240 million nodes and 1.3 billion edges.	Various
7	UC Irvine Machine Learning Repository	https://archive.ics.uci.edu/datasets	668 datasets	Various
8	OpenML	https://www.openml.org/	5731	CC-BY
9	WikiMidea Dumps	https://dumps.wikimedia.org/	NA	Creative Commons Zero (CC0)
10	DataHub	https://datahub.io/collections	11,381	Various
11	HDX - Humanitarian Data Exchange	https://data.humdata.org/dataset	19059	Various
12	Data.Gov	https://catalog.data.gov/dataset	299,941	Various
13	Zenodo	https://zenodo.org/	NA	Various
14	OECD Data Explorer	https://data-explorer.oecd.org/	Data from 2000 regions in 36 countries	CC-BY-4.0
15	UNDATA	https://data.un.org/Default.aspx	32 databases - 60 million records	Various

2.7.6. DATAHUB

Built and owned by LinkedIn, DataHub is an open-source data catalog for the modern data stack. It's a meta-data powered platform. It helps the researchers in data cataloging, data discovery and data governance. In its collections page there are datasets from the arenas such as Air Pollution data, Bibliographic data, Broadband data, Climate Change, Demographics (population), Economic Data and Indicators, Education, Football, GeoJSON, Health Care Data, Inflation, Linked Open Data, Logistics, Machine Learning /Statistical, Movies and TV, Open Corporates, Property Prices, Reference Data, Stock Market Data, War and Peace, Wealth, Income and Inequality, World Bank and YAGO (Yet Another Great Ontology).

Anant Bhardwaj et al have studied Collaborative Data Analytics with DataHub [17] and demonstrated the important aspects of DataHub such as Flexible data storage, native versioning facility, thrift-based data serialization, data analysis in any combination of 20+ languages etc.

Our major aim is to search for IoT datasets, and we applied some search phrases in the various dataset search portals listed in Table I. First of all, we searched for a vague phrase "IoT Attacks" in the 10 dataset search portals and the result we obtained is listed in the Table II. An important point we noted here is Google Data Search and Kaggle gave most appropriate datasets, though the phrase used for search was markedly vague. Quantumstat being a NLP dataset indexing platform, obviously there are no datasets related to IoT. Similarly, Data.gov also didn't give good search results. Therefore, we decided to search Google Dataset Search and Kaggle with different search phrases with respect to Security Issues in IoT.

We searched for "IoT Security", "IoT threats", "Machine Learning in IoT". And then we modified all these search phrases by adding the word Dataset at the end.

Table 2 Datasets listed in GDS and Kaggle for the search phrase "Machine Learning in IoT"

Sl. No	Title/Description	File Type/size	Contributor
1	Temperature Readings: IOT Devices	CSV/1MB	Atul Anand Jha
2	Advanced IoT Agriculture 2024	CSV/3MB	Wisam Abdullah
3	Smart Home Dataset with Weather Information	CSV/20MB	Taranveer Singh Anttal
4	Time Series in IOT (Traffic Data)	CSV/373KB	Vetrivel-PS
5	Edge-IIoTset Cyber Security Dataset of IoT & IIoT	CSV/2GB	Mohamed Amine FERRAG
6	Room Occupancy Detection Data (Temperature, Humidity, Light, CO_2, Humidity Ratio, time)	CSV/269KB	Luis M. Candanedom and Véronique Feldheim
7	Smoke Detection Dataset	CSV/6MB	Stefan Blattmann
8	ACI IoT Network Traffic Dataset 2023 – Dataset generated at IoT Research Lab of Army Cyber Institute, USA	CSV/88GB	Nathaniel Bastian, David Bierbrauer, Morgan McKenzie, Emily Nack
9	IoT Device Network Logs – Data generated from Ultrasonic Sensor with Arduino and NodeMCU	CSV/52MB	Sahil Dixit
10	RT-IOT 2022[36] . Available at https://archive.ics.uci.edu	CSV/3.4MB Proprietary dataset	Sharmila B S Rohini Naga Padma
11	Waste Classification Data	CSV/222MB	Shashank Sekar
12	ICU Healthcare Security Dataset - a use case of an IoT-based ICU with the capacity of 2 beds, where each bed is equipped with nine patient monitoring sensors	CSV/108MB	Faisal Malik
13	Sensor Based Aquaponics Fishpond Datasets	CSV/101MB	Blessing Ogbuokiri, Dr Collins N. Udanor et al
14	Air Temperature of the Greenhouse based on IOT	CSV/20MB	Peyman Daei Rezaei, Beatrice Faniyi, Zhenhua Luo
15	Sensor-Fusion Smoke Detection Classification	CSV/1.47MB	Gaurav Dutta
16	Real-Time Pond Water Dataset for Fish Farming[37]	CSV/13KB	Md. Monirul Islam

17	Customer Transaction Data with IOT Variables – shopping mall dataset with sensor data	CSV/10MB	Ali Ahmed
18	IoT-23: A labelled dataset with malicious and benign IoT network traffic[38]	CSV/ 21.5 GB	Sebastian Garcia Maria Jose Erquiaga Agustin Parmisano
19	Mirai Worm Detection - created from the Canada Institution of Cybersecurity's CICIoT2023 database	CSV/1.22GB	Alexandre Le Mercier
20	NYC FloodSense Sensor Data-Flood Sense Street sign mounted flood depth sensor	CSV/1.74GB	Challagonda,Praneeth;Mydlarz,Charlie;Henaff,Elizabeth;Silverman;Andrea Brain; Tega Khan; Junaid
21	Head gesture recognition with Capacitive Sensors[39]	CSV/2MB	Ionut-Cristian Severin; Dan-Marius Dobrea; Dobrea Monica-Claudia
22	DoS/DDoS-MQTT-IoT - DoS/DDoS attack scenarios in MQTT	CSV/42GB	Alaa Alatram
23	HL-IoT Dataset - high- and low-volume DDoS attack instances of the IoT network[40]	CSV/10.17GB	Makhduma F. Saiyed Irfan Al-Anbagi
24	UFPI NCAD IoT_Attacks - Dataset that consists of IoT attacks based on the MQTT protocol	CSV/275MB	NCAD-UFPI
25	IoT Traffic Generation Patterns Dataset - obtained in the laboratory of Prof. Volkan Rodoplu at Yasar University, Turkey [41]	MATLAB/ 51.96 MB	TUBITAK 118E277, The Scientific and Technological Research Council of Turkey

2.8. IOT DATASETS

As the world is rapidly embracing the IoT based smart technologies in every aspect of human life, the malevolent users blow-out threat wiles to disrupt the smooth working of smart devices and networks. IoT being the network of constrained resources is highly susceptible to security threats. Therefore, all over the world researchers are diligently working to build highly robust defence mechanisms for IoT networks. In this direction, ML and DL models have become helpful to predict or detect the threats. But the machine learning programming concepts need enormous datasets for training purpose. In particular, the classification, prediction and intrusion detection solutions for network security depend on network traffic obtained from real networks. Publicly available datasets are also making scientific experiments more precise and reusable. In this section, we present a comprehensive survey on datasets which contain IoT network traffic[18].

2.8.1. DARPA 98, KDD CUP 98 and KDD CUP 99 Datasets:

The Defence Advanced Research Project Agency, Lincoln Laboratory at MIT and Air Force Research Laboratory together created DARPA dataset to evaluate different IDS available at that time. This is also known as DARPA 98 dataset. The KDD-CUP-98 dataset and the accompanying documentation are created by Ismail Parsa and Ken Howes. The entire folder contains Data dictionary, PKZIP compressed raw learning data set, PKZIP compressed raw validation data set, UNIX compressed raw learning data set, UNIX compressed raw validation data set. In KDD98 there are 481 features and 191779 instances where as KDD99 has 4M instances and 42 features. These sets are available at University of California Irvine Machine Learning Repository and used for The Second and Third International Knowledge Discovery and Data Mining Tools Competitions.

In [19] Stolfo et al applied the algorithms and tools of MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) to process audit data, mine patterns, construct features, and build RIPPER classifiers. In [20] Manoj Kumar Putchala selected the KDD Cup 1999 Intrusion Detection Dataset for the experiments and proposed an intelligent solution which satisfies the key requirements of the IoT solutions and performed the feature engineering using a Random Forest classifier. In [21] M. Tavallaee et al found two vital issues which extremely affects the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, the authors proposed a new data set, NSL-KDD, consisting of selected records of the KDD data set and solved the shortcomings.

2.8.2. BOT-IOT Dataset[22]

BOTNET attack on IoT networks is a well-known issue and several attempts have been made by researchers to create testbeds and capture network traffic with attack scenarios. Creating a realistic IoT network traffic dataset that includes botnet attack scenarios is a challenging task. The BoT-IoT dataset was developed by setting up a realistic network environment in the Cyber Range Lab at University of New South Wales, Canberra, Australia. This environment combined both normal and botnet traffic to provide a comprehensive dataset. We highly recommend the researchers working in the field of Smart Home Networks to use the BOT-IoT dataset because it consists of FIVE IoT scenarios namely a weather station(to record air pressure, humidity and temperature) , A smart fridge (to measure fridge temperature and automatically adjust when it goes below threshold), Motion activated lights , A remotely activated garage door (that opens/closes based on probabilistic input), A smart thermostat (which controls the Air-Conditioner) Attacks are executed by a group of compromised machines called Bots, and target a remote machine, generally a server. The dataset comprises of Botnet attacks DoS, DDoS, Information Theft (Keylogging and Data Theft), and reconnaissance[22]..

The article [22] is the original research work of the creators of the dataset and the dataset has been used by hundreds of scholars to detect and predict DoS, DDoS attacks on IoT and WSN networks. In [23] Jared M. Peterson conducted a feature analysis of BOT-IoT dataset. Sapna Sadhwani et al in [24] used BOT-IoT dataset and developed a lightweight model for DDoS attack detection by Comparing and analysing the performance of machine learning algorithms like LR, NB, RF, ANN, and KNN with binary and multiple-class classification. Mudassir M et al in [25] applied Deep LSTM, GRU, and ANN algorithms on this dataset.

2.8.3. TON-IOT Dataset[3]

In the Cyber Range Labs of UNSW Canberra, Nour Moustafa et al created this Dataset by creating a testbed network for the industry 4.0 network that includes IoT and IIoT devices and services. This new dataset was called TON_IoT because it collected Telemetry data, Operating Systems data and Network data[3].

The BOT-IOT dataset was a simple IoT dataset, and it was generated using a testbed which consisted of only sensors used mainly in smart home networks. Therefore, in the same year, Nour Mustafa et al deployed a new and improvised testbed consisting of many heterogeneous VMs with Windows, Linux, Kali Linux operating Systems interconnecting between the three layers of IoT, Cloud and Edge/Fog systems. A set of sensors, such as green gas IoT and industrial IoT actuators, is connected to MQTT gateways to publish and subscribe to various topics, such as measuring temperature and humidity[3]. We strongly recommend TON-IoT dataset for those researchers who would like to work upon FOG Computing, Edge Computing, IIoT or Industry 4.0 etc. The dataset contains 2,233,921 records of normal and attack data.

In the article [26] Nour Moustafa et al explained the major components of the testbed using which this dataset was created, with respect to Edge layer, Fog layer and Cloud layer. The authors utilized several hacking scenarios and have produced the list of nine attack types viz., Scanning attack, Denial of Service (DoS) attack, Distributed Denial of Service (DDoS) attack, Ransomware attack, Backdoor attack, Injection attack, Cross-site Scripting (XSS) attack, Password attack, Man-In-The-Middle (MITM) attack. This dataset has been used in their studies by authors of [27], [28], [4], [29], [30] and so on. A search in Google Scholar with the phrase “Articles on TON- IoT dataset” gives 1590 search results.

2.8.4. N-BAIoT Dataset [31]

This multivariate dataset is contributed by Yair Meidan et al and the authors proposed the dataset in their article [31]. It is created by collecting *real* traffic data, gathered from 9 commercial IoT devices such as Security Camera, Baby Monitor, Thermostat, Webcam and doorbell that were connected via Wi-Fi to different access points, the devices were connected to a central switch and an router through a wired connection. The authors authentically infected Mirai and BASHLITE botnet attacks and gathered data packets using network sniffer like Wireshark. The dataset consists of 7062606 instances with 23 features[31]. A Google search with the phrase ‘Articles on NBAIoT dataset’ produces About 166 results and thus shows that there are aplenty issues that can be taken up for research works in future.

Jiyeon Kim et al in [**3**2], used the N-BaIoT dataset and developed a botnet detection model using numerous ML models, including deep learning (DL) models. Authors also analysed the effective models with a high detection F1-score by conducting multiclass classification, as well as binary classification, for each model. Teh Boon Seong et al in [33] compared the performances of ML algorithms DT,NB, RF and SVM in purely wireless dataset NaBIOT with traditional wired dataset NSL -KDD. Celil OKUR and Murat DENER in their work [34] carried out study with and without supervision. They used Expectation Maximization algorithm for unsupervised learning and Decision Tree algorithm for supervised algorithm. Priya R. Maidamwar et al in [35] developed a multi-layer perceptron classifier along with Random Forest Algorithm to examine the accuracy, precision, recall and F-Score of IDS. IoT environment-based intrusion related benchmark datasets UNSWNB-15 and N_BaIoT are utilized in the experiment.

In Kaggle, more than 350 IoT datasets are available and many of them are just used in coding and no significant research articles are available which shows that there is an ample opportunity for development of ML and DL models using these datasets. The owners or contributors of the datasets are also seldom bothered to publish papers about their works. The Table 2 lists handful of the major datasets which shall be considered for future research works. Similarly other dataset platforms and search platforms offer thousands of datasets.

3. Conclusion

It's researchers' responsibilty to gather complete information about the datasets before choosing a dataset for their studies. Conducting research work using BOT-IOT dataset to detect Ransomeware attack is not appropriate because the dataset doesn't contain any packets/data with Ransomeware attack information. Instead one can use TON-IoT dataset because this dataset contains rows with Ransomeware attack data. To develop security defense systems or prediction models using ML or DL for WSNs, using datsets like KDD or NSL KDD is not desirable as KDD is not specialized for wireless in general and WSN in particular, even though many researchers have used it to deal with fraud and intrusion detection.

3.1. Research suggestions

The TON-IOT is dataset specially created using heterogeneous environment of devices. The TON-IoT dataset is a collection of datasets that evaluate the efficiency of AI-based cybersecurity applications for the Internet of Things (IoT) and Industrial IoT (IIoT). There for Deep Learning and machine Learning models for edge computing and FOG computing shall be developed using this dataset, Graph Neural Network (GNN) is a DL algorithm which is currently becoming popular. Development of a GNN IDS is a highly recommendable research opportunity for DL enthusiasts. The search using the Phrase "GNN using TON-IoT" in Google Scholar gives only 105 results, which shows that very less works has been done in this field. All DL algorithms can be tested on TON-IoT.

Compliance with ethical standards

Disclosure of conflict of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication

References

[1] A. M. Yavanoglu, Ozlem, "A Review on Cyber Security Datasets for Machine Learning Algorithms," *IEEE international conference on Big Data*, pp. 2186–2193, 2017.

[2] A. Chapman *et al.*, "Dataset search: a survey," *The VLDB Journal*, vol. 29, no. 1, pp. 251–272, Jan. 2020, doi: 10.1007/s00778-019-00564-x.

[3] N. Moustafa, "New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Datasets," 2019.

[4] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and Adna N Anwar, "TON-IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.

[5] N. Moustafa, "New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Datasets," *eResearch Australia Asia 2019*, no. October, pp. 21–22, 2019.

[6] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems : techniques , datasets and challenges," 2019.

[7] R. Kalsotra and S. Arora, "A Comprehensive Survey of Video Datasets for Background Subtraction," *IEEE Access*, vol. 7, pp. 59143–59171, 2019, doi: 10.1109/ACCESS.2019.2914961.

[8] Y. Liu, J. Cao, C. Liu, K. Ding, and L. Jin, "Datasets for Large Language Models: A Comprehensive Survey," Feb. 27, 2024, *arXiv*: arXiv:2402.18041. Accessed: Sep. 09, 2024. [Online]. Available: http://arxiv.org/abs/2402.18041

[9] K. Banachewicz and L. Massaron, *The Kaggle book: data analysis and machine learning for competitive data science*. in Expert insight. Birmingham: Packt Publishing, 2022.

[10] D. Brickley, M. Burgess, and N. Noy, "Google Dataset Search: Building a search engine for datasets in an open Web ecosystem," in *The World Wide Web Conference*, San Francisco CA USA: ACM, May 2019, pp. 1365–1375. doi: 10.1145/3308558.3313685.

[11] O. Benjelloun, S. Chen, and N. Noy, "Google Dataset Search by the Numbers," Jun. 11, 2020, *arXiv*: arXiv:2006.06894. Accessed: Sep. 07, 2024. [Online]. Available: http://arxiv.org/abs/2006.06894

[12] S. A. Hussain, "Google Dataset Search," *American Journalism*, vol. 36, no. 3, pp. 416–417, Jul. 2019, doi: 10.1080/08821127.2019.1644101.

[13] J. N. Van Rijn *et al.*, "OpenML: A Collaborative Science Platform," in *Advanced Information Systems Engineering*, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds., in Lecture Notes in Computer Science, vol. 7908. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 645–649. doi: 10.1007/978-3-642-40994-3_46.

[14] B. Bischl *et al.*, "OpenML Benchmarking Suites," Nov. 22, 2021, *arXiv*: arXiv:1708.03731. Accessed: Sep. 07, 2024. [Online]. Available: http://arxiv.org/abs/1708.03731

[15] G. Casalicchio *et al.*, "OpenML: An R package to connect to the machine learning platform OpenML," *Comput Stat*, vol. 34, no. 3, pp. 977–991, Sep. 2019, doi: 10.1007/s00180-017-0742-2.

[16] B. Bilalli, A. Abelló, and T. Aluja-Banet, "On the predictive power of meta-features in OpenML," *International Journal of Applied Mathematics and Computer Science*, vol. 27, no. 4, pp. 697–712, Dec. 2017, doi: 10.1515/amcs-2017-0048.

[17] A. Bhardwaj *et al.*, "Collaborative data analytics with DataHub," *Proc. VLDB Endow.*, vol. 8, no. 12, pp. 1916–1919, Aug. 2015, doi: 10.14778/2824032.2824100.

[18] F. De Keersmaeker, Y. Cao, G. K. Ndonda, and R. Sadre, "A Survey of Public IoT Datasets for Network Security Research," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 3, pp. 1808–1840, 2023, doi: 10.1109/COMST.2023.3288942.

[19] W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 227–261, Nov. 2000, doi: 10.1145/382912.382914.

[20] M. K. Putchala, "Deep Learning Approach for Intrusion Detection System (IDS) in the Internet of Things (IoT) Network using Gated Recurrent Neural Networks (GRU)," *Wright State University*, p. 64, 2017.

[21] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, ON, Canada: IEEE, Jul. 2009, pp. 1–6. doi: 10.1109/CISDA.2009.5356528.

[22] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019, doi: 10.1016/j.future.2019.05.041.

[23] J. M. Peterson, "A REVIEW AND ANALYSIS OF BOT-IOT SECURITY DATA FOR MACHINE LEARNING," *MACHINE LEARNING*.

[24] S. Sadhwani, B. Manibalan, R. Muthalagu, and P. Pawar, "A Lightweight Model for DDoS Attack Detection Using Machine Learning Techniques," *Applied Sciences*, vol. 13, no. 17, p. 9937, Sep. 2023, doi: 10.3390/app13179937.

[25] M. Mudassir, D. Unal, M. Hammoudeh, and F. Azzedin, "Detection of Botnet Attacks against Industrial Systems by Multilayer Deep Learning Approaches," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–12, May 2022, doi: 10.1155/2022/2845446.

[26] N. Moustafa, M. Keshk, E. Debie, and H. Janicke, "Federated TON_IoT Windows Datasets for Evaluating AI-based Security Applications," Oct. 04, 2020, *arXiv*: arXiv:2010.08522. Accessed: Sep. 08, 2024. [Online]. Available: http://arxiv.org/abs/2010.08522

[27] N. Moustafa, E. Adi, B. Turnbull, and J. Hu, "A New Threat Intelligence Scheme for Safeguarding Industry 4.0 Systems," *IEEE Access*, vol. 6, pp. 32910–32924, 2018, doi: 10.1109/ACCESS.2018.2844794.

[28] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. H. D. Hartog, "ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 485–496, Jan. 2022, doi: 10.1109/JIOT.2021.3085194.

[29] A. R. Gad, M. Haggag, A. A. Nashat, and T. M. Barakat, "A Distributed Intrusion Detection System using Machine Learning for IoT based on ToN-IoT Dataset," *IJACSA*, vol. 13, no. 6, 2022, doi: 10.14569/IJACSA.2022.0130667.

[30] [30] S. Soni, M. A. Remli, K. M. Daud, and J. Al Amien, "Performance evaluation of multiclass classification models for ToN-IoT network device datasets," *IJEECS*, vol. 35, no. 1, p. 485, Jul. 2024, doi: 10.11591/ijeecs.v35.i1.pp485-493.

[31] Y. Meidan *et al.*, "N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018, doi: 10.1109/MPRV.2018.03367731.

[32] J. Kim, M. Shim, S. Hong, Y. Shin, and E. Choi, "Intelligent Detection of IoT Botnets Using Machine Learning and Deep Learning," *Applied Sciences*, vol. 10, no. 19, p. 7009, Oct. 2020, doi: 10.3390/app10197009.

[33] T. B. Seong, V. Ponnusamy, N. Zaman Jhanjhi, R. Annur, and M. N. Talib, "A comparative analysis on traditional wired datasets and the need for wireless datasets for IoT wireless intrusion detection," *IJEECS*, vol. 22, no. 2, p. 1165, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1165-1176.

[34] C. Okur and M. Dener, "Detecting IoT Botnet Attacks Using Machine Learning Methods," in *2020 International Conference on Information Security and Cryptology (ISCTURKEY)*, Ankara, Turkey: IEEE, Dec. 2020, pp. 31–37. doi: 10.1109/ISCTURKEY51113.2020.9307994.

[35] Department of Computer Science & Engineering, G H Raisoni University, Amravati, India, 444701, P. R. Maidamwar, P. P. Lokulwar, and K. Kumar, "Ensemble Learning Approach for Classification of Network Intrusion Detection in IoT Environment," *IJCNIS*, vol. 15, no. 3, pp. 30–36, Jun. 2013, doi: 10.5815/ijcnis.2023.03.03.

[36] B. S. Sharmila and R. Nagapadma, "Quantized autoencoder (QAE) intrusion detection system for anomaly detection in resource-constrained IoT devices using RT-IoT2022 dataset," *Cybersecurity*, vol. 6, no. 1, p. 41, Sep. 2023, doi: 10.1186/s42400-023-00178-5.

[37] Md. M. Islam, M. A. Kashem, and J. Uddin, "Fish survival prediction in an aquatic environment using random forest model," *IJ-AI*, vol. 10, no. 3, p. 614, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp614-622.

[38] S. Garcia, A. Parmisano, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic." Zenodo, https://zenodo.org/records/4743746, Jan. 20, 2020. doi: 10.5281/zenodo.4743746.

[39] I.-C. Severin and D.-M. Dobrea, "Head Gesture Recognition Based on Capacitive Sensors Using Deep Learning Algorithms," *Bulletin of the Polytechnic Institute of Iași. Electrical Engineering, Power Engineering, Electronics Section*, vol. 67, no. 3, pp. 73–92, Sep. 2021, doi: 10.2478/bipie-2021-0018.

[40] M. F. Saiyedand and I. Al-Anbagi, "Deep Ensemble Learning With Pruning for DDoS Attack Detection in IoT Networks," *Trans. Mach. Learn. Comm. Netw.*, vol. 2, pp. 596–616, 2024, doi: 10.1109/TMLCN.2024.3395419.

[41] TUBITAK 118E277, "IoT Traffic Generation Patterns Dataset." Kaggle, Jan. 2021. [MATLAB]. Available: https://www.kaggle.com/datasets/tubitak1001118e277/iot-traffic-generation-patterns