ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Enhancing IoT Security: A Machine Learning Approach for Spam Detection

[1]KOYAGURI SINDHU, [2]MUCHARALA CHERRISMAA, [3]CH Y S V S GOPAL, [4]Mrs.K. SOWJANYA NAIDU, [5]Dr.K.VASANTH KUMAR, [6]Mr.L.RAMU

Department of Computer Science and Engineering- Internet of Things
Malla Reddy Engineering College, Hyderabad, Telangana.
cherrismaa@gmail.com

Abstract— **The Internet of Things (IoT) comprises millions of sensor and actuator-equipped devices connected via wired or wireless channels for data transmission. With over 25 billion devices projected to be interconnected by 2020, the volume of data generated is expected to grow significantly. Machine learning algorithms can enhance IoT system security and usability by detecting anomalies and ensuring authentication based on biometric data. However, attackers may exploit vulnerabilities in IoT systems. To address this, we propose a machine learning-based approach to detect spam in IoT devices, evaluating five models against various input feature sets to compute a spam score reflecting device trustworthiness. Using the REFIT Smart Home dataset, our approach demonstrates superior effectiveness compared to existing methods.**

I. INTRODUCTION

The Internet of Things (IoT) facilitates connectivity and integration among real-world objects regardless of their locations, posing significant challenges for privacy and security. Protecting data privacy in IoT applications is crucial to mitigate security threats such as intrusions, spoofing, DoS attacks, jamming, eavesdropping, spam, and malware [1].

Security measures for IoT devices vary based on organization size, type, and user behavior, necessitating cooperation among security gateways. Location, nature, and application of IoT devices dictate security measures, as exemplified by smart security cameras in organizations. Web-dependent IoT devices, common in workplaces, require careful handling to prevent security breaches [2].

Approximately 25-30% of employees connect personal IoT devices to organizational networks, exposing them to potential security risks. The evolving IoT landscape attracts both users and attackers, prompting the adoption of defensive strategies by IoT devices, leveraging machine learning to optimize security protocols while balancing security, privacy, and computational constraints [3]. However, implementing effective security protocols remains challenging due to resource limitations and the dynamic nature of IoT networks and attack scenarios.

A. Contributions:

1) Validation of spam detection scheme using five distinct machine learning models.
2) Proposal of an algorithm for computing spamicity scores to facilitate detection and decision-making.
3) Analysis of IoT device reliability based on computed spamicity scores using various evaluation metrics.

B. Organization:

The paper proceeds as follows: Section II reviews related work, Section III presents the proposed scheme, Section IV discusses and analyzes results, and Section V concludes the paper II.

II. NARRATIVE REVIEW

IoT systems face vulnerabilities from network, physical, and application attacks, along with privacy breaches involving objects, services, and networks, as depicted in Fig. 1. Here are some examples of attack scenarios initiated by malicious actors.

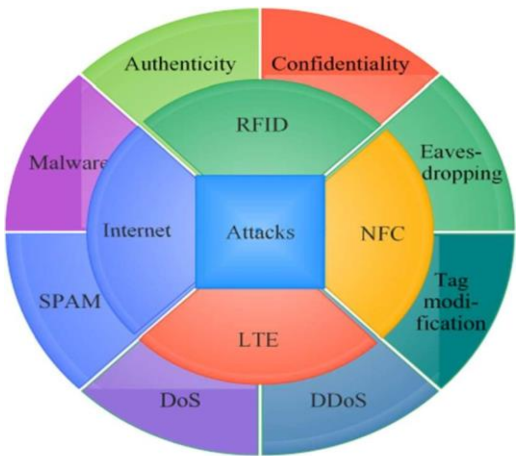


Fig. 1: Protocols with possible attacks

• DDoS attacks flood target databases with unwanted requests, blocking access to services. These attacks, orchestrated by IoT botnets, exhaust service provider resources, rendering networks unavailable [3].

• RFID attacks target IoT devices physically, compromising device integrity by tampering with data storage or transmission.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Common attacks include availability, authenticity, and confidentiality breaches, countered by measures like password protection and data encryption [4].

• Internet attacks involve spammers using techniques like Ad fraud to generate artificial clicks for profit, disrupting targeted websites. Cybercriminals exploit unencrypted traffic and tag modification in NFC attacks, countered by conditional privacy protection and random public keys [5][6].

• Supervised machine learning techniques such as SVMs, random forest, and neural networks detect attacks like DoS, DDoS, intrusion, and malware in IoT devices [7][8][9][10].

• Unsupervised machine learning techniques like multivariate correlation analysis detect DoS attacks in IoT by forming clusters without labels [11].

• Reinforcement machine learning techniques like Q-learning improve authentication and malware detection by allowing IoT systems to select security protocols and key parameters [12][9][13].

ML enables lightweight access control protocols, extending IoT system lifetimes. K-NNs are applied to address unregulated outer detection in WSNs, enhancing network security [14].

Machine learning techniques, as demonstrated in literature, play a vital role in detecting web spam, offering a diverse range of approaches for implementation [15].

III. PROPOSED SCHEME

A. System model:

The modern digital landscape heavily relies on smart devices, necessitating spam-free information retrieval from them. Gathering information from diverse IoT devices poses a significant challenge due to the multitude of domains involved, resulting in the generation of vast amounts of heterogeneous and varied data, termed as IoT data. This data is characterized by features like real-time updates, multiple sources, and a mix of rich and sparse content.

B. Proposed methodology:

The effectiveness of managing IoT data improves when stored, processed, and retrieved efficiently.
This proposal seeks to minimize spam occurrence from these devices, as indicated by Eq.

$$\min P(s) = \aleph - \sim s \quad \textbf{(1)}$$

In Eq. 1, ℵ represents the information collection. ~s denotes the vector of spam-related information, which is subtracted from ℵ to reduce the likelihood of receiving spam information from IoT devices.

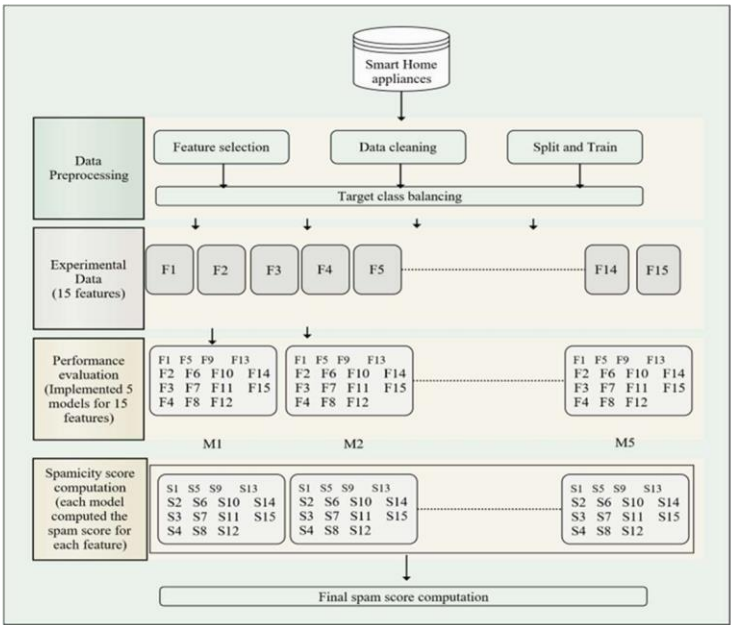


Fig. 2: Approach followed in the proposed scheme

Targeting Web Spam Detection for IoT Devices

To safeguard IoT devices from generating malicious information, this proposal focuses on web spam detection. Various machine learning algorithms are considered for spam detection from IoT devices, particularly targeting issues within home deployments. The proposed methodology meticulously addresses data engineering parameters before validation with machine learning models.

1) Feature Engineering:

Feature reduction aims to decrease data dimensionality, addressing issues like overfitting and resource requirements. Principal Component Analysis (PCA) is a popular technique for feature extraction [15]. In this proposal, PCA is combined with IoT parameters, such as analysis time and web-based appliance usage, to streamline feature extraction effectively.

• Analysis Time:

Data from an eighteen-month span is condensed to one month to enhance accuracy, considering months with maximum climate variations.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Author	Machine learning technique	Target attack	Performance
Kulkarni et al., 2009 [7]	Neural Network	DOS	Improved the performance of system
Tan et al., 2013 [11]	Multivariate correlation analysis	DOS	Improved accuracy
Li et al. ,2016 [12]	Q-Learning	DOS	Solved the associated optimality equations
Alsheikh et al., 2014 [8]	SVM, Naive Bayes	Intrusion	Detected the WSN attacks successfully
Buczak et al., 2015 [9]	Machine learning techniques	Cyber attacks	survey of ML techniques for detection of cyber attacks
Xiao et al.,2017 [13]	Q-Learning	Malware	Improve the detection accuracy
Narudin et al., 2016 [10]	Random forest, K-NN	Malware	99.97% true-positive rate (TPR)

TABLE I: Machine learning techniques used for the detection of different attacks.

• Web-Based Appliances:
Only appliances reliant on web connectivity are included in data collection, ensuring relevance to IoT device functionality.

2) Feature Selection:
Entropy-based filtering, utilizing correlation among discrete and continuous attributes, determines feature importance [17]. Functions like information.gain and symmetrical.uncertainty assess feature relevance based on training data attributes.

Model no.	Model	Method	Package	Tuning parameters
Model1	Bagged Model	Bag	Caret	Vars
Model2	Bayesian Generalized Linear Model	bayesglm	Arm	None
Model3	Boosted Linear Model	BstLm	bst, plyr	mstop, nu
Model4	eXtreme Gradient Boosting	xgbLinear	Xgboost	nrounds, lambda, alpha
Model5	Generalized Linear Model with Stepwise Feature Selection	glm-StepAIC	MASS	None

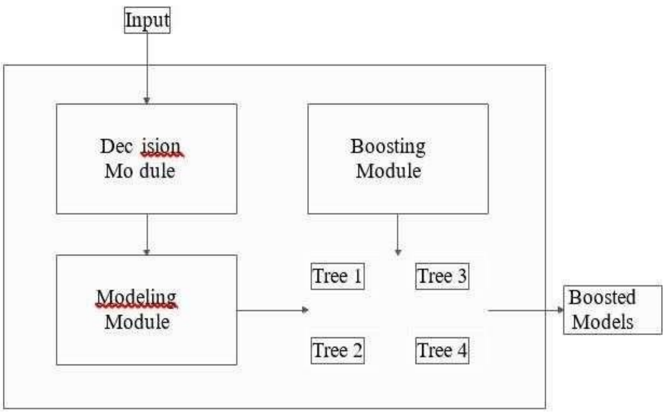


Fig. 3: Boosted linear model phases.

C. Machine Learning Models:

1) Bayesian Generalized Linear Model (BGLM): BGLM is a log-likelihood uni-modal for exponential family forms, emphasizing essential Bayesian elements [18][19].

• Incorporates prior information quantitatively specified as a distribution representing coefficient probability.

• Pairs prior with a likelihood function, resulting in probability function outcomes. Combination of prior and probability function forms subsequent coefficient value distributions.

• Simulations from posterior distribution construct empirical population parameter value distributions.

• Simple statistics summarize posterior distribution and simulate statistical distribution.

2) Boosted Linear Model: Creates multiple decision trees for data elements, modeling each data group as a linear function. Boosted models are formed from these modeling modules [20].

Algorithm 1 Spamicity score computation

Input:
Output: Computed spamicity score

```
procedure FUNCTION(PageRank)
    for i = 1 to n do
        for j = 1 to 15 do
            Matrix representation z_i                . Formulation of matrix: n*15
            Set j ← j + 1
            Set i ← i + 1
        end for
    end for
    for i = 1 to 15 do
        Set V_i =← x        . Where x is the feature importance score according to
  Table III                 . chine Learning model building
    end for
    p[i] ← Y                . Where Y is the predicted constraint
for i = 1 to 15 do
        Compute RMSE[i]= sqrt( Σ_{i=1}^{n} (p_i − a_i)^2 / n )   ▷ p_i is the predicted array and
  a_i is the actual array
    end for
        for i = 1 to 15 do S ← RMSE[i] * V_i
    end for
end procedure
```

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

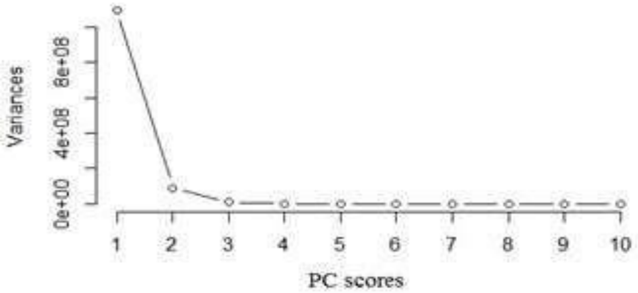
Vol 18, Issue 2, 2024

• Built iteratively in each training round, adjusting parameters to minimize prediction errors.

• Utilizes gradient calculations to adjust system parameters and minimize errors in subsequent learning rounds [21].

4) Generalized Linear Model with Stepwise Feature Selection: Generalized Linear Models (GLMs) provide a versatile framework for interpreting dependent variables using multiple predictor variables [22]. Stepwise feature selection is employed to identify significant effects in the equation, iteratively repeating until all significant effects are found.

D. Spamicity Score: After evaluating machine learning models, spamicity scores for each appliance are computed to indicate device trustworthiness and reliability [23]. Spamicity score computation involves attribute importance scores and error rates, as defined by Eq. 2 and Algorithm 1 implemented in R.



E. Complexity Analysis: The algorithm's complexity is assessed by evaluating all steps and their respective iterations.

Time Complexity: Steps 2 to 8 involve linear matrix formulations, requiring O(n) time.

$$e[i] = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

$$S \leftarrow RMSE[i] * V_i$$

• In the worst-case scenario, loops in steps 2-8, 9-11, and 13-15 also take O(n) time.

• Steps 10, 12, and 14 have constant-time calculations, resulting in O(1) time complexity.

• Time complexity (TC) is calculated as follows:

[TC = O(n) + O(n) + O(n) + O(1)]

Feature	attr importance
plugIdRef	0.76342
spaceIdRef	0.12322
manufacturer	0.23432
model	0.20345
Occupancy Type	0.10346
builtFormType	0.20998
wallAgeBand	0.43219
conditionType	0.76908
roomType	0.03076
wallType	0.38151
windowType	0.12602
fuelType	0.06642
meterType	0.47700
Heading	0.30532
Battery.Life	0.61396

TABLE IV: Summary of performance of the experimental models

Model	Precision	Recall	Accuracy	Score distribution
M1	0.650	1	79.81	Refer Fig. 5
M2	0.541	1	83.22	Refer Fig. 6
M3	0.567	1	84.35	Refer Fig. 7
M4	0.598	1	88.9	Refer Fig. 8
M5	0.513	1	91.8	Refer Fig. 9

Space Complexity: The algorithm's space complexity is determined by assessing memory usage.

- Input size not exceeding (n) contributes to (O(n)) space complexity.
- Loops also contribute (O(n)) space.
- Arithmetic operations take (O(1)) space.
- Space complexity is calculated as : [SC = O(n) + O(n) + O(1)]

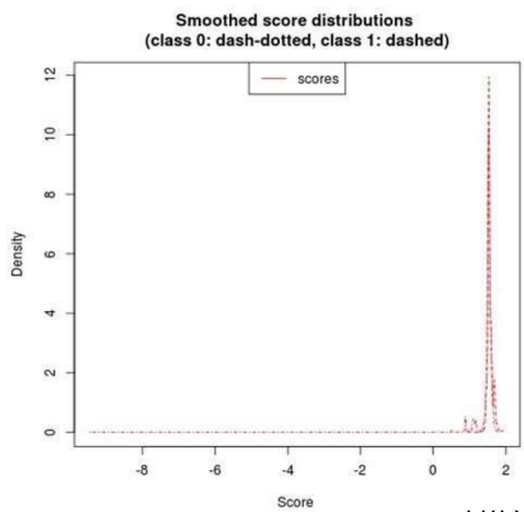


Fig. 5: Spam score distribution by Bayesian Generalized Linear Model

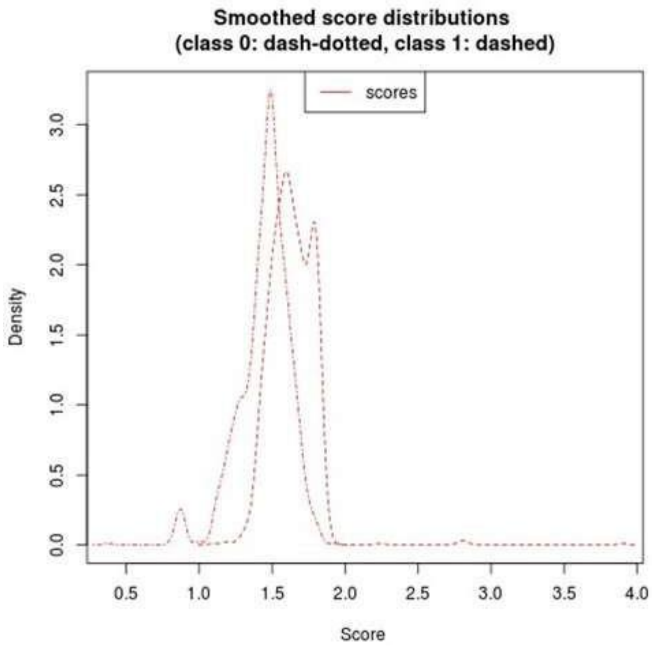


Fig. 6: Spam score distribution by Bagged Model

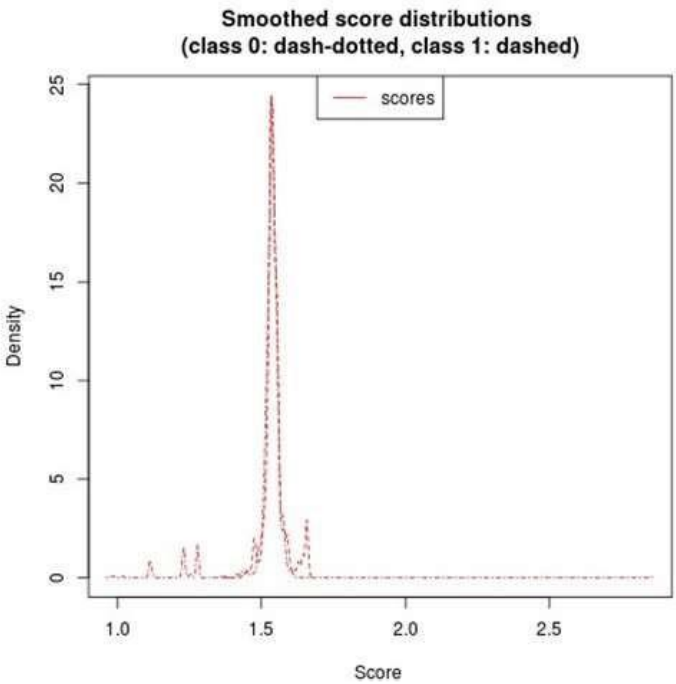


Fig. 7: Spam score distribution by Boosted Linear Model

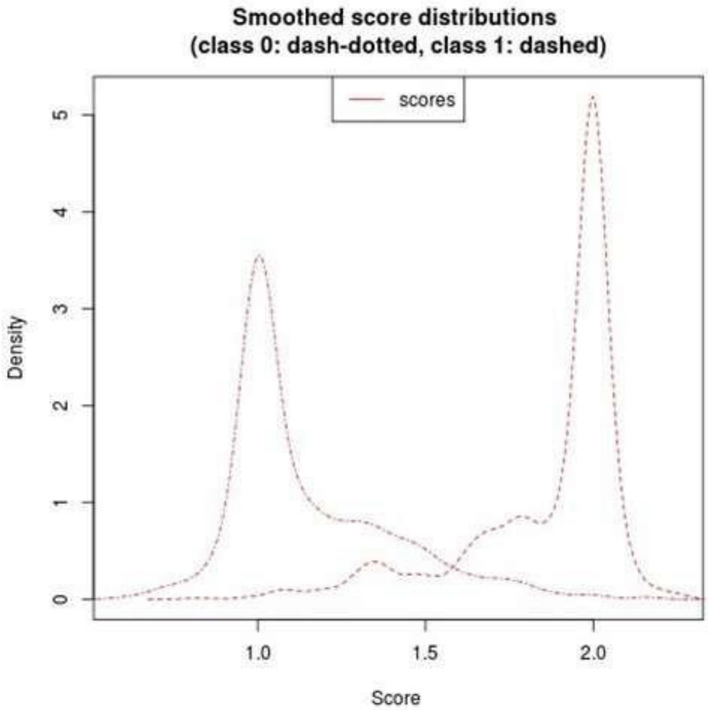


Fig. 8: Spam score distribution by eXtreme Gradient Boosting

Fig. 8: Spam score distribution by eXtreme Gradient Boosting.

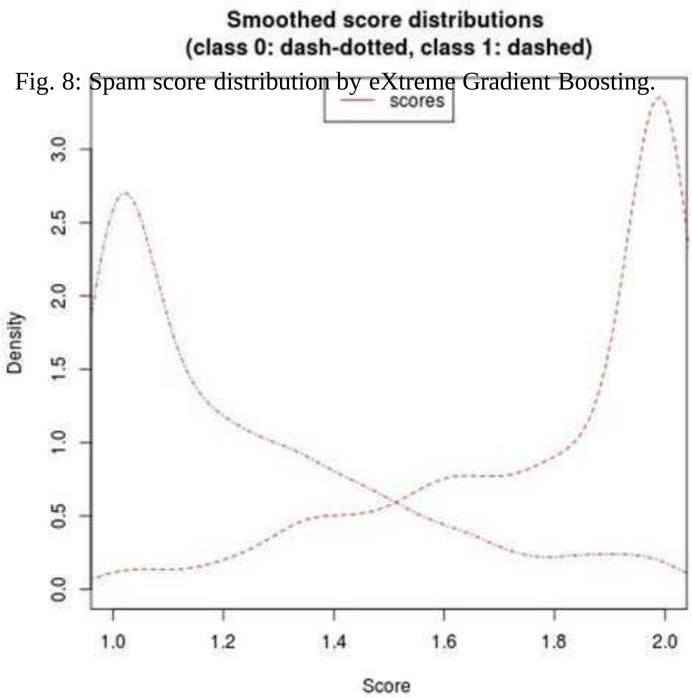


Fig. 9: Spam score distribution by Generalized Linear Model with Stepwise Feature Selection.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

TABLE V: Spamicity score of appliances

Appliance	Internet Connectivity√	M1	M2	M3	M4	M5
Air filter		0.65	0.396	0.399	0.371	0.628
Alarm clock	×	0.348	0.580	0.947	0.637	0.2168
Alarm radio	×	0.246	0.607	0.686	0.633	0.175
Aquarium	×√	0.671	0.709	0.143	0.878	0.489
Baby monitor		0.734	0.701	0.625	0.216	0.651
Bread maker	×	0.820	0.683	0.261	0.789	0.217
CD player	×√	0.066	0.657	0.369	0.782	0.220
Chiller		0.045	0.635	0.466	0.732	0.213
Coffee grinder	×	0.081	0.283	0.046	0.074	0.020
Coffee maker	×	0.138	0.6150	0.312	0.210	0.562
DAB radio	×	0.092	0.234	0.554	0.773	0.208
Dehumidifier	×√	0.160	0.106	0.608	0.761	0.223
Desktop PC		0.981	0.615	0.558	0.8188	0.274
Dishwasher	×√	0.691	0.6090	0.542	0.16	0.230
Docking station		0.135	0.206	0.602	0.881	0.235
Doorbuster	×√	0.186	0.613	0.631	0.905	0.228
DVD player/recorder		0.204	0.610	0.625	0.944	0.897
Electric blanket	×	0.244	0.009	0.648	0.008	0.219
Electric heater	×	0.012	0.006	0.011	0.012	0.220
Electric toothbrush charger	×	0	0	0	0	0
Exercise machine	×	0.341	0.211	0.132	0.429	0.227
Fairy lights	×√	0.402	0.578	0.062	0.921	0.230
Games console	√	0.453	0.563	0.825	0.9620	0.240
George Forman grill	√	0.486	0.558	0.840	0.985	0.235
Guitar amplifier		0.477	0.558	0.795	0.928	0.229
Hair tongs	×√	0.507	0.5548	0.840	0.470	0.2306
Hifi	√	0.556	0.548	0.865	0.938	0.838
iPad/iPod docking station	√	0.593	0.423	0.892	0.992	0.2319
Kitchenette	√	0.621	0.535	0.917	0.987	0.230
Laptop		0.633	0.534	0.925	0.964	0.928
Lava Lamp	×√	0.617	0.538	0.227	0.285	0.224
Microwave	√	0.637	0.531	0.938	0.933	0.225
Oven	√	0.647	0.529	0.789	0.937	0.227
PC monitor	√	0.657	0.529	0.955	0.949	0.226
Printer	√	0.667	0.528	20.798	0.946	0.227
Projector		0.367	0.926	0.960	0.959	0.892

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

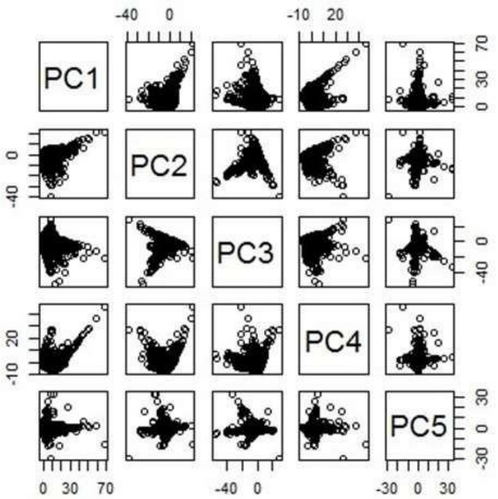
Vol 18, Issue 2, 2024

Radio	×√	0.686	0.525	0.344	0.610	0.229
Raspberry Pi		0.686	0.243	0.966	0.973	0.886
Scanner	×√	0.695	0.230	0.110	0.212	0.228
Router	√	0.523	0.975	0.974	0.874	0.751
Record player	√	0.963	0.981	0.977	0.911	0.2291
Set top box		0.177	0.473	0.735	0.754	0.7520
Sewing machine	×	0.542	0.509	0.199	0.921	0.221
Shredder	×	0.606	0.572	0.721	0.196	0.541
Tape player	×	0.231	0.806	0.738	0.701	0.684
Telephone	×√	0.770	0.739	0.751	0.707	0.005
Television		0.718	0.751	0.743	0.712	0.779
Toaster	×√	0.105	0.211	0.657	0.123	0.231
Washing machine		0.729	0.725	0.809	0.778	0.992

Feature	PC1	PC2	PC3	PC4	PC5 ——	PC15
1	4.255091e-08	6.764816e-05	1.145414e-06	-4.126413e-07	2.332671e-04	-1.612771e-12
2	1.257375e-04	-1.348555e-04	3.608422e-12	7.430535e-12	1.237237e-12	1.480848e-04
3	4.948566e-11	4.266645e-03	-1.223795e-12	1.007857e-02	9.111890e-12	4.042344e-05
4	7.535564e-04	4.896944e-02	1.090096e-02	9.787808e-03	1.816266e-01	4.702625e-02
5	2.637138e-01	4.681924e-02	9.005530e-13	5.998283e-01	6.321595e-02	-2.265900e-14
6	1.620736e-01	8.626930e-15	2.347263e-01	6.220893e-01	-1.063215e-01	4.663576e-16
7	6.879058e-01	2.180458e-01	-2.698411e-01	3.001963e-01	-4.495283e-15	5.822666e-01
8	-9.253025e-02	-6.857088e-01	2.269870e-03	6.833382e-01	1.559366e-04	-1.408902e-01
9	6.522830e-01	-1.762764e-03	6.512795e-01	4.664394e-02	-3.218220e-01	-1.127804e-15
10	-2.196190e-02	2.877072e-05	-2.021959e-15	-1.085136e-03	-1.139319e-05	4.868426e-05
11	3.189077e-12	2.859939e-05	1.615075e-04	-1.230201e-11	-7.347401e-05	1.216977e-11
12	6.950183e-05	3.858547e-12	1.745346e-04	1.637610e-02	1.778308e-10	1.681497e-13
13	8.558204e-10	-6.920804e-07	4.439540e-14	-8.191861e-06	2.146017e-12	-5.372825e-05
14	-2.058280e-15	-3.574847e-03	1.067373e-9	5.693648e-05	-4.831610e-02	-1.984294e-09
15	6.29293e-07	3.15414e-09	5.92394e-07	-1.23342e-07	-4.15506e-07	9.95639e-07

TABLE VI: Principal components being computed by PCA method for features.

IV. RESULTS AND DISCUSSION



A. Data Collection

• A smart home dataset was collected by the REFIT project [20] sponsored by Loughborough University.

• The dataset encompasses sensor data from 20 homes, capturing internal environmental conditions for 18 months.

• Each home included over 100,000 data points collected across various rooms

• This openly available dataset can be found at [20].

B. Experimental Setup:

• Data traces from the REFIT project dataset [20] were used for the experiments.

• RStudio, an open-source software (available at [21]), was employed for analysis.

• The software requires Windows 7/8/10, macOS 10.12+,

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Ubuntu 14/16/18, or Debian 8/10.

C. Impact of Data Preprocessing on SDI-UML:

Preprocessing involved selecting appliances to identify spam parameters.

- Feature reduction using Principal Component Analysis (PCA) aimed to decrease data dimensionality.
- PCA generates principal components (PCs) corresponding to each data point.
- In this dataset with 15 features, 15 PCs were obtained.

D. Impact of Machine Learning Models on SDI-UML

- Five machine learning models were trained using the features from Table III.
- Each model generates a "spamicity score" for each appliance, indicating its susceptibility to spam.
- Table IV summarizes the performance of these models.
- Table V lists the selected appliances with their corresponding spamicity scores.
- Figures 5-9 depict the distribution of spamicity scores across the models.
- Model evaluation metrics include accuracy, precision, and recall (details omitted).

V. CONCLUSION

Leveraging machine learning, this framework detects spam affecting IoT devices. The approach utilizes a pre-processed IoT dataset for training various machine learning models. These models assign a "spam score" to each appliance, enabling better control over smart home functionality. Future work aims to incorporate climatic and surrounding features to enhance IoT device security and trustworthiness.

VI. REFERENCES

[1] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, "Iot security: ongoing challenges and research opportunities," in 2014 IEEE 7th international conference on service-oriented computing and applications. IEEE, 2014, pp. 230–234.

[2] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for iot security and privacy: The case study of a smart home," in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.

[3] E. Bertino and N. Islam, "Botnets and internet of things security," Computer, no. 2, pp. 76–79, 2017.

[4] Annemneedi Lakshmanarao, Ampalam Srisaila, Tummala Srinivasa Ravi Kiran, Kamathamu Vasanth Kumar, Chandra Sekhar Koppireddy, "An efficient smart grid stability prediction system based on machine learning and deep learning fusion model", Indonesian Journal of Electrical Engineering and Computer Science, Vol.33, No.2,February2024,pp. 1293~1301,http://doi.org/10.11591/ijeecs.v33.i2.pp1293-1301.

[5] W. Kim, O.-R. Jeong, C. Kim, and J. So, "The dark side of the internet: Attacks, costs and responses," Information systems, vol. 36, no. 3, pp. 675–705, 2011.

[6] H. Eun, H. Lee, and H. Oh, "Conditional privacy preserving security protocol for nfc applications," IEEE Transactions on Consumer Electronics, vol. 59, no. 1, pp. 153–160, 2013.

[7] R. V. Kulkarni and G. K. Venayagamoorthy, "Neural network based secure media access control protocol for wireless sensor networks," in 2009 International Joint Conference on Neural Networks. IEEE, 2009, pp. 1680–1687.

[8] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996– 2018, 2014.

[9] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2015.

[10] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," Soft Computing, vol. 20, no. 1, pp. 343–357, 2016.

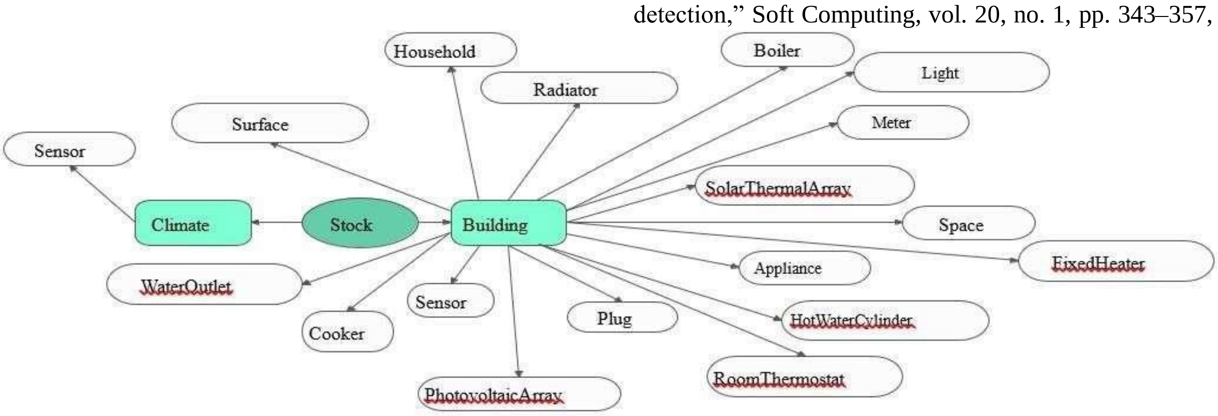


Fig. 11: Features of Smart Home dataset