International Research Journal of Multidisciplinary Scope (IRJMS), 2024; 5(1):74-85

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2024.v05i01.0146

Machine Learning Driven Spam Detection for IoT Devices

Blessa Binolin Pepsi M*, Renuka S, Rajeshwari K, Abithanjalee PV

Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India *Corresponding Author's Email: mblessa@mepcoeng.ac.in

Abstract

Smart homes are gradually incorporating Internet of Things (IoT) devices, producing massive amounts of data. Wireless methods are used to convey the vast majority of this data. Several IoT devices may be impacted by a variety of IoT risks, including cyberattacks, inconsistent network connectivity, data leakage, etc. Machine learning (ML) techniques could be quite helpful in this case to guarantee security and authorization. Data anomalies must be discovered using statistics. This study investigates the dependability of IoT equipment that interacts. When determining a spam score, the six Machine Learning models are taken into account with increased input features including xgboost, BGLM, GLM, Bagging, and Stacking. The grade demonstrates the dependability of an IoT device based on many criteria. This grade demonstrates the dependability of an IoT device based on many criteria. When compared to other recent approaches, the findings demonstrate the effectiveness of the proposed strategy. The spamicity score of the home IoT device is calculated using the Spam Score algorithm. A publicly available dataset for smart homes meteorological data from the UCI repository and synthetic data from IoT devices are used to validate the technique. The results collected show just how effective the suggested algorithm is in analyzing time series data from the Internet of Things devices for spam identification.

Keywords: Classification, Feature Engineering, Internet of Things, Machine Learning, Regression, Spam Detection.

Introduction

An interconnected network of distributed embedded systems that communicate with one another using wired or wireless communication methods is known as the Internet of Things (IoT) (1). The huge size of IoT device deployment makes it possible to access a wide range of data produced by these devices, and modeling and analysing the data for general anomalies is a critical challenge (2). The security measures for IoT devices are determined by their location, nature, and use (3-4). Smart homes and smart cities now contain IoT devices owing to the Internet of Things (IoT) rapid growth and rapid proliferation. A network of physical objects or things with restricted computing, storage, and communication capabilities that are also integrated with electronics (like sensors and actuators), software, and network connectivity that enable them to gather process, and share data is also referred to as a "thing network". Numerous industries, including vital infrastructure, agriculture, the military, home appliances, and personal health care, are served by the Internet of Things. IoT devices adopt a defensive strategy and determine the essential parameters in the security protocols for trade-offs between security, privacy, and computing, albeit, with the introduction of machine learning (ML) in various attack scenarios

The number of abnormalities brought on by IoT devices is increasing along with the number of these devices (5). IoT objects include smart home appliances like smart bulbs, smart adapters, smart meters, smart refrigerators, smart ovens, air conditioners, temperature sensors, smoke detectors, and IP cameras as well as more advanced gadgets like frequency identification (RFID) devices, heartbeat detectors, accelerometers, parking zone sensors, and various other sensors in cars, among others. These devices are exposed to network attacks such as data theft, phishing, spoofing, and denial of service attacks (DDoS attacks) (6). These may trigger further cyber security risks like ransomware assaults and significant data breaches, which may cost firms a lot of money and time to repair. In general, sensors experience a wide range of problems caused by faulty hardware, short battery life or human error. There are some ways to detect faulty sensors in

(Received 19th October 2023; Accepted 05th January 2024; Published 30th January 2024)

an IoT environment (7-9) by using the correlation either between activities and sensors between actuators and sensors. Typically, it is difficult for an IoT system with limited resources to estimate the current network and timely attack state, making this job complex. According to market research, 50% of employees who are now at work connect their own IoT devices to the company network. Both consumers and attackers are drawn to the IoT because of its increasing nature.

In the proposed computation of spamicity algorithm, the spamicity score is determined. The spam level or percentage of any appliances, websites, emails, etc. is referred to as the "spamicity score." The spamicity score for the REFIT dataset (10) has been determined here. There are several household appliances in this dataset.

This framework primarily focuses on the reliability of IoT devices. We are unable to frequently check for spam attacks on all household appliances. Here, spam is computed using the individuals' current voltage consumptions. We can determine whether it has spam or not by looking at the current flow statistics. Machine learning algorithms are frequently used by hackers to pinpoint the weaknesses in IoT smart devices. Both consumers and attackers are drawn to the IoT because of its increasing nature. IoT devices adopt a defensive strategy and determine the essential parameters in the security protocols for tradeoffs between security, privacy, and computing, albeit, with the introduction of machine learning (ML) in various attack scenarios. In the same way that spam networks made up of infected PCs are referred to as "botnets," they have evolved into "thingbots," which Proofpoint describes as commandeered IoT devices.

The remainder of this paper is structured as follows. The relevant work was described in Section II. The suggested plan is illustrated in Section III. The results are discussed and analyzed in Section IV. Section V completes this article's conclusion. Context extraction was offered as a method for the automatic detection of malfunctioning Internet of Things (IoT) devices. In two stages, the process works. The system precomputes sensor correlation and the chance that one sensor state will vary from another, or context, during a precomputation phase. The system discovers a break of sensor correlation and transition during a real-time phase to find and identify the defects. Here presented a DICE (11), a context-based, future-oriented strategy to identify malfunctioning IoT devices for real-time datasets. SMOTE method focuses on the data anomalies that are common in smart Internet of Things (IoT) devices. Dealing with models that were trained on unbalanced data is difficult. Therefore, we suggest using ensemble learning along with the Synthetic Minority Over-Sampling Technique. SMOTE (12) is known for its wider usage in anomaly detection as it is flexible in terms of data size. SMOTE enables the development of normal classifiers from imbalanced data sets that are representative of real-world data. Network attackers are currently targeting interconnected systems like Web servers, database servers, and cloud computing servers. Denial-of-service (DoS) attacks have a serious impact on these computing systems. DoS attack detection system employs multivariate correlation analysis (MCA) (13) to determine the geometrical connections between network traffic elements for reliable network traffic categorization (14). This technique is used to differentiate between known and unknown DoS attacks from legitimate network traffic.

Monitoring, identifying issues, and treating industrial processes are all done by IoT systems in the industry. The SaaS provisioning may be used by the majority of manufacturing systems to perform MFDD operations. Scalability, semantic support, space efficiency, consolidation schemes, and other issues are overlooked to improve the framework design to address them. MFDM SaaS offers a set of privacy- preserving mining services to ensure the security policies of each tenant are strictly observed. The information in the repository that is private to the tenants may be publicly sharable, sharable after sanitization, or not sharable. Enhance the overall experience in a smart home setting by maximizing user adaptation, identifying issues in lifestyle, raising alerts, enhancing reminder systems, and assisting prompting systems (15).

Although blockchain-based systems offer outsourced security and anonymity, they come with a high energy, latency, and processing overhead, making them unsuitable for the majority of IoT devices with limited resources.

The smart organization's IoT security cameras, for instance, can record many parameters for analysis and wise decision-making (16). Due to its democratized, secure and, private nature, Blockchain technology empowers Bitcoin, the first cryptocurrency system. Here evaluating the efficiency of our defence against two serious security breaches that are particularly pertinent to smart homes. The first one is a distributed denial of service (DDOS) assault, in which the attacker overwhelms a specific target node by using a number of infected IoT devices. IoT devices have been exploited in a number of recent incidents to launch powerful DDoS attacks. The second is a linking attack, where the attacker creates a connection between several transactions or data ledgers with the same PK to discover. By breaking into a number of smart home devices, such as a light bulb, switch, and smoke alarm, authors showed that off-the-shelf IoT devices are lacking in fundamental security measures. They said that even if the home gateway manages the packet exchange to and from the home, smart homes are still susceptible to attacks launched from users' smartphone. IoT-based Botnet attacks are proof that IoT device subsidies are still relevant as long as they function properly and safely. However, when these tools are misused or do not produce consistent results and securely, both their benefits and drawbacks become apparent (17). Machine learning algorithms have defeated this attack. A lot of different types of assaults can occasionally be found in the network data. By examining the network traffic data sets and carrying out a learning process, this unfavourable circumstance can be made favourable. The data set includes attacks by IoT devices and malicious network traffic. Only one of the trials analysed in this research had feature reduction done. The study's objective is to use machine learning to accurately discern between legitimate traffic and attack traffic in a network. In this paper, two separate learning models were used, however in our paper, just one learning model was used and it provided the accuracy and efficient results that were desired.

Methods for choosing wrapper features are meant to decrease the number of features in the original feature set while increasing classification accuracy simultaneously. By reducing the redundant, irrelevant, and noisy data from the original dataset, FS-Feature Selection aims to improve the accuracy (18). FS is viewed as a binary optimisation issue, with solutions limited to the binary values "0" and "1" (19). Therefore, it is appropriate to apply the binary form of the DA method to resolve this issue. The result of the problem is represented as a vector of “0” and “1”, where a zero means that the relevant feature is not selected and a one indicates that this feature is selected. In this (20) work, a rapid filter approach that can quickly find relevant features and redundancy among relevant features without pairwise correlation analysis and also introduce a novel concept called predominant correlation. It will be considered a good feature for the classification task if the correlation between a feature and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other relevant features. This approach demonstrates its efficiency and effectiveness in dealing with high dimensional data for classification. Wrapper or embedded methods improve predictor performance (21). The idea of Internet of Things (IoT) is implanting networked heterogeneous detectors into our daily life. It opens extra channels for information submission and remote control to our physical world. A significant feature of an IoT network is that it collects data from network edges. Moreover, human involvement for network and devices maintenance is greatly reduced, which suggests an IoT network need to be highly self-managed and self-secured. For the reason that the use of IoT is growing in many important fields, the security issues of IoT need to be properly addressed. Among all, Distributed Denial of Service (DDoS) is one of the most notorious attacking behaviours over network which interrupt and block genuine user requests by flooding the host server with huge number of requests using a group of zombie computers via geographically distributed internet connections. (22) DDoS disrupts service by creating network congestion and disabling normal functions of network components, which is even more disruptive for IoT. In this paper, a lightweight defensive algorithm for DDoS attack over IoT network environment is proposed and tested against several scenarios to dissect the interactive communication among different types of network nodes (23).

In this proposed work, Blockchain-based infrastructure is proposed to support security- and privacy-oriented spatio-temporal smart contract services for the sustainable Internet of Things (IoT)-enabled sharing economy in mega smart cities. The infrastructure leverages cognitive fog nodes at the edge to host and process off loaded geo-tagged multimedia payload and transactions from a mobile edge and IoT nodes, uses AI for processing and extracting significant event information, produces semantic digital analytics, and saves results in Blockchain and decentralized cloud repositories to facilitate sharing economy services. The framework offers a sustainable incentive mechanism, which can potentially support secure smart city services, such as sharing economy, smart contracts, and cyber-physical interaction with Blockchain and IoT (24).We explore and provide taxonomies of the causes and costs of the attacks, and types of responses to the attacks.

Methodology

Spam Detection with ML

Utilizing cutting-edge technologies offers individuals, groups, and states incredible benefits, some people have been prejudiced against them. The security of locations to store extremely sensitive data, information portability, and other issues are a few instances. Given these issues, one of the greatest issues we currently have is digital oppression driven by dread. Due to a number of groups, including the criminal underworld, professionals, and digital advocates, including the fear of the digital world, which has caused many issues for individuals and organizations, has reached a point where it could jeopardize national and open security. Intrusion Detection Systems (IDS) were created as a consequence to safeguard against online assaults. Because it is gathered from various areas, information retrieval from various IoT devices is a significant challenge. IoT generates a large amount of heterogeneous, diverse data due to the involvement of numerous devices. This type of information can be referred to as IoT data. Real-time, multi-source, complex, and sparse are just a few of the characteristics of IoT data. By feature reduction and selection, the data are classified and the spamicity score has been calculated.

Feature Engineering

This method is used to reduce the dimensionality of the day. In this project, we have used the PCA (25) method for feature reduction. PCA is one of the mostly used algorithms in Machine Learning projects to reduce the data dimensions. The essential idea when using PCA as a method for feature selection is to pick variables depending on the magnitude (from biggest to least in absolute values) of their coefficients. PCA rotates data from one coordinate system to another. By using this approach, over-fitting, memory requirements, and computing capacity issues are reduced. For feature separation, there are several techniques. The data collected over a period of eighteen months is included in the dataset utilized in the studies. We took into account data from one month in order to get better outcomes and accuracy. There were now 15 characteristics left in the data dimensions. This method is also like feature reduction. But this is used for pre-processing the data. In our work, Entropy Based Filter algorithm is used for testing. This is used for better reduction in our dataset and used for pre-processing. To determine the weights of discrete qualities, this technique analyses the correlation between discrete and continuous attributes. Based on their entropy with a continuous class attribute, rate the relevance of discrete qualities. The information in F-Selector is reimplemented in this method. Gain, Gain Rate, and Symmetric Uncertainty. The entropy-based filtering technique, which also discloses users' hidden interests, aids in the solution of the cold-start problem. In trials, three different collaborative filtering recommendation systems are compared for accuracy using real-world data from MAE metrics. The results show that the entropy-based algorithm (26) provides superior suggestion quality than the user-based method and achieves recommendation accuracy comparable to the item-based strategy.

The Figure 1 depicts a high-level overview of the system architecture of the Spam Detection. The input data is collected from a dataset on smart homes, and it has been pre-processed to deal with missing values and remove unnecessary columns.

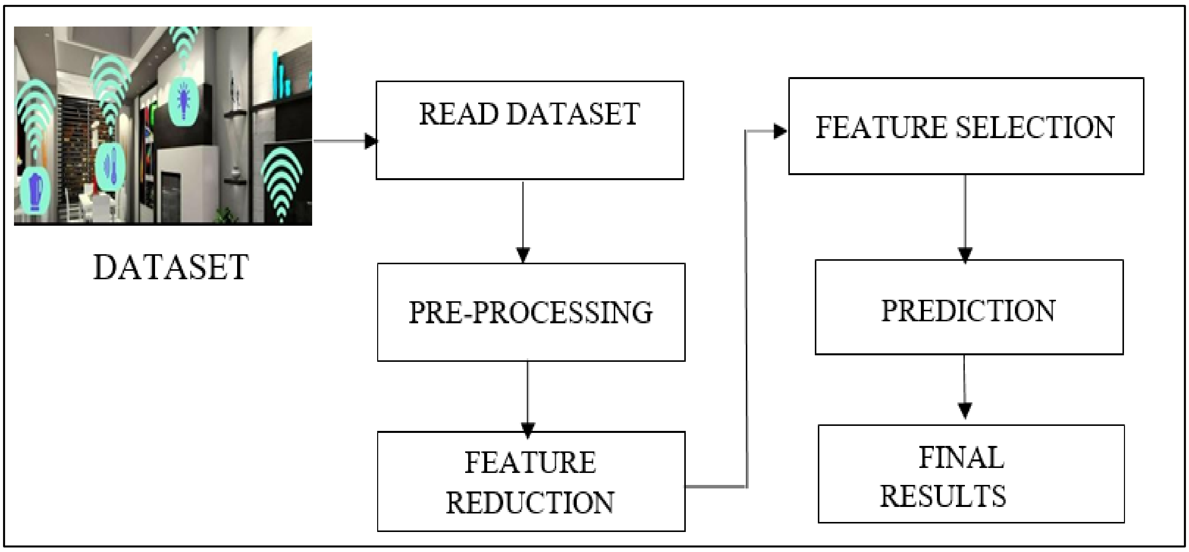


Figure 1: Architecture Diagram of the proposed work

By utilising the principal component analysis approach and a quick correlation-based filter, the dataset's characteristics are condensed to fifteen in total. To improve performance, the data is divided into train and test groups. The model is categorised and the spamicity score method is used once it has been categorised using machine learning techniques like bagging, boosting, etc. The findings are then compared and demonstrated.

Machine Learning Models

Five distinct Machine Learning models are used in this research (27-30). The performance of exponentially growing IoT data requires backing up, parsing, and successfully being recovered. The goal of this proposition is to decrease the amount of spam originating from the devices described in Eq1.They are separately calculated using the spamicity score technique, and finally, the spam likelihood is determined using the RMSE scores values.

$P(s) = N - \vec{s}$ [1]

In Eq.1, N refers to the gathering of data. To reduce the likelihood of receiving spam information from IoT devices, $\vec{S}$ is the vector of spam-related information that is removed.The Entropy Based Filter algorithm approach examines the association between discrete and continuous characteristics to calculate the weights of discrete features. Determine the importance of discrete attributes by evaluating their entropy using a continuous class property. This approach reimplements the data in F-Selector, Symmetric Uncertainty, Gain, and Gain Rate. The users' hidden interests are also revealed by the entropy-based filtering method, which helps solve the cold-start issue. Using actual data from Mean Absolute Error (MAE) metrics, three different collaborative filtering recommendation systems are tested and their accuracy is evaluated. The findings demonstrate that the entropy-based algorithm achieves recommendation accuracy equivalent to the item-based approach and offers better suggestion quality than the user-based method.

Bayesian Generalized Linear Model (BGLM):

For exponential family forms, it is a consistent, asymptotically effective, and asymptotically normal log likelihood uni-modal. The actual focus of Bayesian approaches is on these fundamental components. The Bayesian technique allows one to account for model complexity even when the model parameters are constrained by the data since applying a prior distribution to the parameters regularizes the fitting procedure.

Generalized Linear Model

A dynamic framework offered by GLMs may be utilized to interpret a dependent variable using a number of explanatory (predictor) variables. The parameter dependent might be either continuous or discrete, and the explanatory variables can be either empirical (covariate) or categorical

(Factors). Using a stepwise feature selection procedure, the model was fitted.

The three components of a GLM (31) are a linear predictor, a link operation, and a distribution of likelihood. The linear predictor represents the linear combination of predictor variables, whereas the link function turns the linear forecast to the scale of the response variable. The probability distribution links the linear predictor's variability to the outcome variable's unpredictability through the link factor.

eXtreme Gradient Boosting (xgboost)

It is an effective and scalable gradient boosting technique. An efficient linear model solver and a tree learning technique are also included in the package. Regression, grouping, and ranking are just a few of the objective functions it offers.

Due to its effectiveness, scalability, and accuracy, xgboost has become more popular in a variety of data science applications and contests. It has a wide variety of hyperparameters that may be adjusted to improve the model's performance and can handle huge datasets with high-dimensional features.

Boosted Linear Model

Multiple decision trees are generated for the data pieces, and the decision tree models are created by categorizing the data series into various data classes. Consequently, each of the data groupings is modeled as a linear function.

Each weak learner in a boosted linear model is a linear regression model, a straightforward linear equation that forecasts the value of a dependent variable based on the values of one or more independent variables. The initial linear model in the boosting technique is fitted to the training data, and subsequent models are added in a way that minimizes the total error of the model.

A weighted total of the predictions from each individual linear model make up the output of a boosted linear model. Each model's weight is a reflection of how crucial it is for producing reliable forecasts. For regression issues where the objective is to forecast a continuous variable or classification challenges where the objective is to predict a categorical variable, boosted linear models can be utilized.

Bagging Model

A method known as bagging, often referred to as bootstrap aggregating, can be used to improve the accuracy and effectiveness of machine learning systems. It is employed to control a prediction model's bias-variance and lower its variance.The bootstrap samples are subsequently taught separately and concurrently using weak or base learners. Finally, a mean or a majority of the predictions are picked, depending on the assignment, to calculate a more precise estimate. Regression involves averaging every output predicted by each classifier; this process is referred to as soft voting. Hard voting, often known as majority voting, is the process of accepting the class that receives the greatest number of votes in classification issues.

Stacking Model

Using stack cords, a mechanism known as stacking (32) joins many switches to create a logical switch for packet transmission. As a widely used horizontal emulation method, it may boost bandwidth, expand the number of terminals, and enhance networking economy. The goals of bagging, boosting, and stacking are to reduce variance, reduce bias, and increase prediction accuracy. Boosting and bagging combine similar weak students. Different solid learners are combined by stacking. Models are progressively trained by boosting and simultaneously trained by bagging. Stacking provides an excellent accuracy rate in this dataset.

Spamicity Score

The spamicity score is obtained at the end of training in all six models. We may assess the trustworthiness of the devices using this score.

$$\mathrm{e(i)} = \sqrt{\frac{\sum_{i=1}^{n}(pi-ai)^2}{n}} \qquad [2]$$

$$X \leftarrow RMSE[i] * Bi \qquad [3]$$

The error rate calculated using the expected and actual arrays is represented by e (i) in the equations above. With the aid of the attribute importance score and mistake rate, the spamicity score, or S, is calculated.

Spamicity Score Detection

Algorithm: Computation of Spam Score

Input: Feature extracted output from the appliances.

Output: Computed spamicity score

```
procedure FUNCTION (SpamScore)
  for i = 1 to 15 do
    Set $b_i$ = ← k
  end for
  p(i) ← Y
  for i = 1 to 15 do
    Compute RMSE (i) = $\sqrt{\frac{\sum_{i=1}^{n}(pi-ai)^2}{n}}$
  end for
  for i = 1 to 15 do
    X ← RMSE (i) ∗$b_i$
  end for
end procedure
```

Table 1: Dataset Description

Name of the dataset	Number of Instances	Number of Features
Smart Home dataset with weather information	50,000	32
Open Smart Home IoT	10,000	2
Smart Home dataset	60,000	10

Table 2: RMSE Values for Smart home dataset with weather information

ML Models	MSE	MAE	RMSE
xgboost	0.04638	0.20798	0.21533
BGLM	0.01836	0.07845	0.13550
BLM	0.00792	0.49588	0.08903
GLM	0.01836	0.07846	0.13550
Bagging	1.74008	0.00020	0.00417
Stacking	1.58359	0.37867	0.00285

Table 3: Performance of 6 Machine Learning Models

Model	Precision	Recall	Accuracy
xgboost	0.68	1	0.57
BGLM	0.59	1	0.33
BLM	0.57	1	0.52
GLM	0.53	1	0.33
Bagging	0.56	1	0.92
Stacking	0.69	1	0.99

Results and Discussion

Experiments

Using ThingSpeak and NodeJS, the performance of the categorization results is evaluated and contrasted with other machine learning models and IoT devices.

Dataset Description

The experimental study is tested on 3 datasets for training the models and 1 dataset for IoT devices. The datasets are Smart home dataset, smart home dataset with weather information, Open smart home IoT datasets. The datasets specification is shown in Table 1. All the dataset mentioned contains only numerical values. Smart home dataset with weather information includes all numerical values and it is represented in kw unit. These datasets are pre-processed, and trained using the following five models xgboost, BGLM, GLM, BLM, and Bagging. From this using the algorithm the spamicity score is evaluated. Spamicity Score is computed that indicates the spam causing probability in the appliances. For IoT part the dataset used is the temperatures that stored in the NodeJS server is used and ThingSpeak is used to analyse the spam and not spam and display a graphical difference.

Performance Metrics

The planned work's efficiency and performance are evaluated using RMSE's parameters of comparison.The Table 2 represents the Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) Scores for all the ML models. Eq4 defines MAE, it gives the absolute value of actual target variable and predicted target value from the observation. Here is how the Eq 5 is defined This equation explains the MSE.

$$\text{MAE} = |y_i - \hat{y}_i| \quad [4]$$

$$\text{MSE} = \frac{\sum_{i=1}^{n}(pi-ai)^2}{n} \quad [5]$$

Figure 2 represents the accuracy of the five machine learning models on the dataset of smart home appliances. The RMSE should be as low as possible. We discovered that bagging has a low value and produces a high accuracy value when compared to other models such as Xgboost, BGLM, BLM, and GLM.From this using the algorithm the spamicity score is evaluated. Spamicity Score is computed that indicates the spam causing probability in the appliances. Table 3 compares the accuracy for various models. The staking model provides the most accuracy while comparing with other machine learning models since it combines both bagging and boosting. For Realtime applications the IoT devices are the temperature of the room is monitored and updated in the cloud platform ThingSpeak, it is an IoT analytics platform where we can visualize and analyse the live data streams. Visualization of Real and Spam temperatures in ThingSpeak is shown in Figure 3 and Figure 4. By classifying the data as either spam or non-spam, the dataset is prepared for machine learning. Manual labelling or supervised learning methods can be used to do this task. Choose a machine learning model that can effectively identify spam in IoT devices from the available models. These can include more sophisticated models like deep learning models as well as more conventional models like decision trees or logistic regression. To detect and stop spam, integrate the trained model with IoT devices. There are several ways to accomplish this, either by employing a cloud-based service or a local edge device.

In Arduino board, a temperature sensor, a Wi-Fi module, 5V-Relay and connection wires are connected. Connect the temperature sensor to the Arduino board by connecting the v_{cc} pin of the sensor to the 5v pin of the Arduino, the GND pin of the sensor to the GND pin of the Arduino, and the signal pin of the sensor to an analog pin of the Arduino. Then, Connect the Wi-Fi module to the Arduino board by connecting the VCC and GND pins of the module to the respective pins of the Arduino and the RX and TX pins of the module to digital pins of the Arduino. To send the data from Arduino to the Node MCU, TX and RX pin of the Node MCU is connected to the fifth and sixth pin in the Arduino. It will transfer sensor data to Node MCU via serial communication. The ground pin in Node MCU is connected to the GND pin in the Arduino. The 5V pin is connected to the V_{in} pin in the Node MCU. The code should read the temperature from the sensor using the analog, read function and then send the data to a remote server using the Wi-Fi module. The server should analyse the data for spam detection. For example, if the temperature reading is outside of a certain range, it could be flagged as spam. The temperature values sense by the sensor is displayed in LCD connected with Arduino and it is shown in Figure 5.

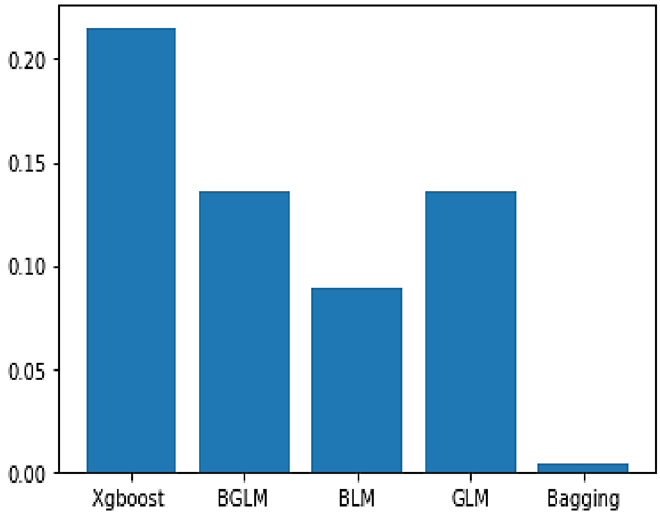


Figure 2: Comparison Graph on various models based on RMSE values

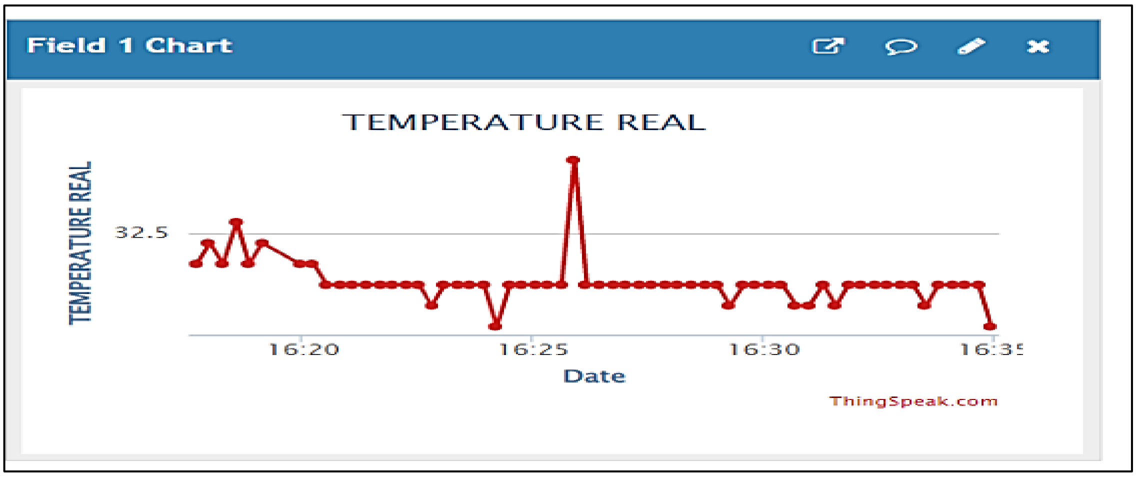


Figure 3: Actual Temperature

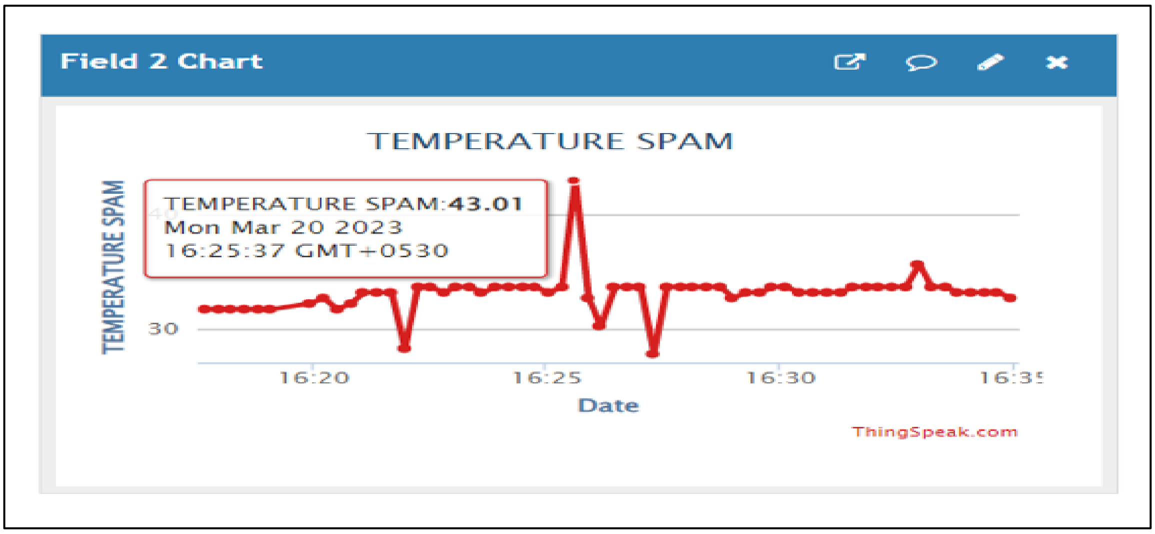


Figure 4: Spam Temperature

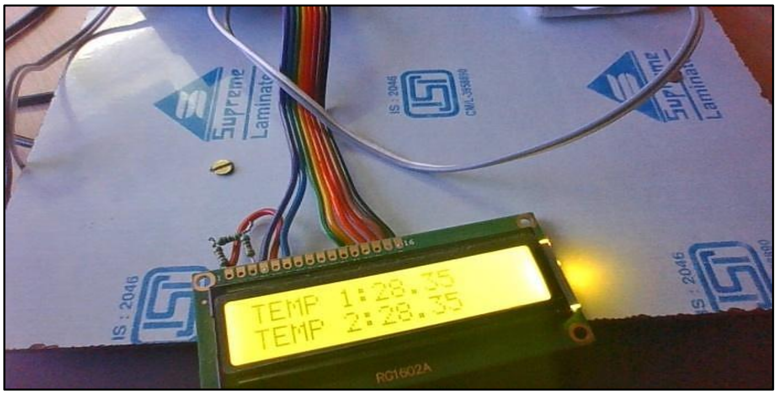
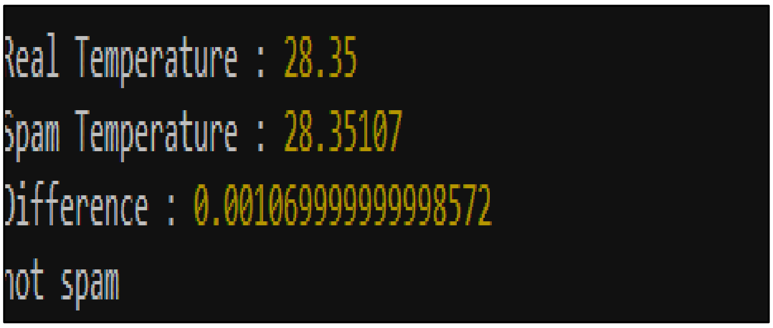
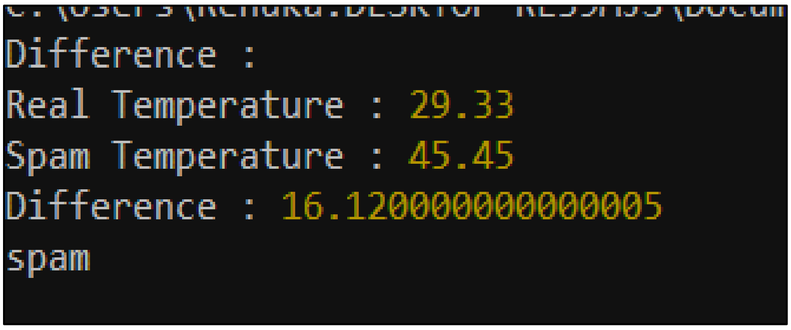


Figure 5: Temperature display in LCD



```
Real Temperature : 28.35
Spam Temperature : 28.35107
Difference : 0.0010699999999998572
not spam
```

Figure 6: Output from NodeJS as NotSpam



```
Difference :
Real Temperature : 29.33
Spam Temperature : 45.45
Difference : 16.120000000000005
spam
```

Figure 7: Output from NodeJS as Spam

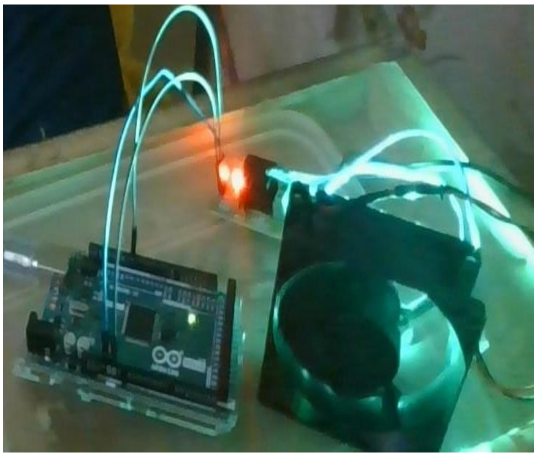


Figure 8: IoT devices connected with Arduino is Turned On

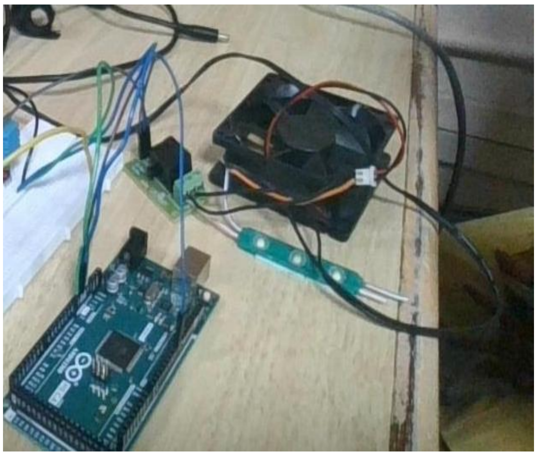


Figure 9: IoT devices connected with Arduino is Turned Off

The node server calculating the spam values from the real and fake temperatures. It provides values that include both the real temperature and the spam temperature. By seeking these two values, it evaluates the distinction and determines whether or not the message is spam. The sample values are presented in Figure 6 is represented as Non spam and Figure 7 is represented as Spam.

When there is no spam is calculated by the server then the home appliances connected with Arduino will be high. If any temperature readings on the server are flagged as spam, then the Appliances will be turned off automatically.

Spam detection experiment will depend on the accuracy of the temperature sensors. The temperatures are updated to the ThingSpeak and analysed for any spam or anomalies in data. If there are sudden changes in the temperature data then it could decide that the data is spam. Figure 8 and Figure 9 shows the changes occurred while temperature rises. This method provides spam detection on IoT devices for home appliances using Arduino and a temperature sensor.

Conclusion

The proposed stacking method uses machine learning (ML) models to identify the spam properties of IoT devices. The feature engineering approach is utilized to pre-process the IoT data collection that was used in the trials. In an experiment, IoT equipment is given a spam score using a framework using ML models. This improves the prerequisites needed for IoT devices in a smart home to operate effectively. The large-scale, diverse objects and networks are the major characteristics that set IoT security concerns apart from traditional ones. IoT security is significantly more challenging to manage as a result of the two elements, heterogeneity, and complexity in synthetic data. It can be extended to more complicated applications and customized further to incorporate further sensors or algorithms for more sophisticated spam detection methods.

Abbreviations

Nil

Acknowledgments

The authors acknowledge the magnanimous support and continuous encouragement by Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India.

Author's contribution

All the authors have contributed to complete this research work.

Disclosure statement

No potential conflict of interest and competing interests are reported by the author(s).

Ethical statement

The paper reflects the author's research and analysis truthfully and completely.

Funding

No funds, grants, or other support is received for conducting this study.

References

1. Sonare B, Dharmale GJ,Renapure A, Khandelwal H, and Narharshettiwar S. E-mail Spam Detection Using Machine Learning. 4th International Conference for Emerging Technology, Belgaum, India. 2023;1-5.
2. Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener. Comput Syst. 2013;29:7.
3. Zhang ZKCho, MCY, Wang CW, Hsu CW, Chen CK, and Shieh S. IoT security: Ongoing challenges and research opportunities. Proc. IEEE 7th Int Conf Service-Oriented Comput Appl. 2014; 230–234.
4. Zhang ZK, Cho MCY, Wang CW, Hsu CW, Chen CK, and Shieh S. IoT security: Ongoing challenges and research opportunities. IEEE 7th Int Conf Service-Oriented Comput Appl. 2021;230–234.
5. Yogendra Singh Parihar, Scientist D and District Informatics Office, National Informatics Centre. Internet Of Things and Nodemcu. 2019;6:6.
6. Zhang C, and Green R. Communication security in Internet of Thing: Preventive measure and avoid DDOS attack over IOT network. 18th Symp Commun Network. 2020;8–15.
7. Ye J, Stevenson G, and Dobson S. Detecting abnormal events on binary sensors in smart home environments. Pervasive and Mobile Computing. 2016;33:32–49.
8. Sikder K, Aksu H, and Uluagac AS. 6thsense: A context-aware sensor-based attack detector for smart devices. USENIX Security Symposium. 2017.
9. Choi J, Jeoung H, Kim J, Ko Y, Jung W, Kim H, Kim J. Detecting and Identifying Faulty IoT Devices in Smart Home with Context Extraction. IEEE/IFIP International Conference on Dependable Systems and Networks. 2018.
10. “REFIT smart home dataset.” [Online] Available:https://repository.lboro.ac.uk/articles/REFIT_Smart_Home_dataset/2070091.2019

11. Isra'a AbdulNabi, Qussai Yaseen, Spam Email Detection Using Deep Learning Techniques. Procedia Computer Science. 2021; 184:853-858.
12. Tang S, Gu Z, Yang Q, Fu S. Smart Home IoT Anomaly Detection based on Ensemble Model Learning from Heterogeneous Data. IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA. 2019; 4185–4190.
13. Zhiyuan Tan, ArunaJamdagni, XiangjianHe, Priyadarsi Nanda. A System for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis. IEEE Transactions on Parallel And Distributed Systems. 2014; 25:2.
14. Zhang C and Green R. Communication security in Internet of Thing: Preventive measure and avoid DDOS attack over IOT network. 18th Symp Commun Network. 2015;15.
15. Jakkula VR, Cook DJ. Detecting Anomalous Sensor Events in Smart Home Data for Enhancing the Living Experience. Artificial intelligence and smarter living. 2011.
16. Rahman MA. Blockchain and IoT-based cognitive edge framework for sharing economy services in a smart city. IEEE Access. 2019; 7:18611–18621.
17. Bertino E, and Islam N. Botnets and Internet of Things security. Computer. 2017;50:76–79.
18. Majdi M. Mafarja, DerarEleyan, Iyad Jaber, Seyedali Mirjalili , Binary Dragonfly Algorithm for Feature Selection. International Conference on New Trends in Computing Sciences (ICTCS). 2017.
19. Abdelaziz I. Hammouri a, MajdiMafarja b, Mohammed Azmi Al-Betar c, Mohammed A. Awadallah d, Iyad Abu-Doush. An improved Dragonfly Algorithm for feature selection Knowledge-Based Systems. 2020;203.
20. Yu L, and Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. Int. Conf. Machine Learning. 2003:856–863.
21. Guyon and A. Elisseeff. An introduction to variable and feature selection. J. Mach. Learn. Res. 2003; 3:1157–1182.
22. Narudin FA, Feizollah A, Anuar NB, and Gani A. Evaluation of machine learning classifiers for mobile malware detection. Soft Computing. 2016;20:343–357.
23. Zhang C and Green R. Communication security in Internet of Thing: Preventive measure and avoid DDOS attack over IOT network. 18th Symp. Commun. Network. 2020; 8–15.
24. Kim W, Jeong OR, Kim C, and So J. The dark side of the internet: Attacks, costs and responses. Inf Syst. 2021;36:675–705.
25. Rahul Adhao, Vinod Pachghare. Feature selection using principal component analysis and genetic algorithm. Discrete Mathematics and Cryptography. 2022;23:2.
26. Davood Gharavian, Mansour Sheikhan, Alireza Nazerieh, Sahar Garoucy. Speech emotion spam detection using FCBF feature selection method and GA-optimized neural network. Neural Computing and Applications. 2012;21:2115 -2126.
27. Yogendra Singh Parihar, Scientist D and District Informatics Office, National Informatics Centre. Internet of Things and Nodemcu. 2019; 6:6.
28. Niranjani V, Agalya Y, Charunandhini K, Gayathri K, and Gayathri R. Spam Detection for Social Media Networks Using Machine Learning. 8th International Conference on Advanced Computing and Communication Systems, Coimbatore, India, 2022; 2082-2088.
29. Omar Saad, Ashraf Darwish, Ramadan Faraj. A survey of machine learning techniques for Spam filtering. IJCSNS International Journal of Computer Science and Network Security. 2012; 12:2.
30. Saini KG and Sharma S. Machine Learning Approaches for an Automatic Email Spam Detection. International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1*)*. 2023;1-5.
31. Xiangming Meng, Sheng Wu, Jiang Zhu. A Unified Bayesian Inference Framework for Generalized Linear Model. IEEE Signal Processing Letters. 2018; 25:3.
32. Bohdan Pavlyshenko. Using Stacking Approaches for Machine Learning Models. IEEE Second International Conference on Data Stream Mining & Processing. 2018.