REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

Optimizing Feature Selection for IIoT Security: A Sequential Machine Learning Strategy

S Gouri Kiran Kumar[1], Dr Raavi Satya Prasad[2],

[1]Research Scholar, Department of Computer Science and Engineering, Acharya Nagarjuna University, Andhra Pradesh, India
kiran43427@gmail.com, gouri.kiran@utas.edu.om
[2]Professor and Dean R & D, Department of Computer Science & Engineering, Dhanekula Institute of Engineering & Technology, Ganguru, Vijayawada, A.P., India, deanresearch@diet.ac.in; orcid: 0009-0007-1894-2417.

Abstract:
IIoT attacks have the potential to stop industrial processes, which can result in lost production, broken equipment, and large financial losses. IIoT systems frequently manage sensitive data, including operational, proprietary, and employee personal data. Attacks may result in information theft, misuse, and data breaches. To depict the behaviors and interactions of IIoT dev ices, extract pertinent elements from the raw data. features may consist of command sequences, resource use measurements, and device communication patterns. Most of the existing systems uses tree-based techniques or embedded techniques like LASSO to further hone the feature set in accordance with model performance. It may not be ideal if all correlated features are significant because LASSO has a tendency to randomly choose one feature from a set of highly correlated data and ignore the others. The proposed model designs a sequence of ml models namely Random Forest followed by Integrated KNN and logistic regression as feature selector because this order guarantees that the features are assessed using several viewpoints: ensemble techniques (Random Forest), non-parametric methods based on distance (KNN), and linear models (Logistic Regression). The balancing of complexity, performance, and interpretability can be accomplished by using these techniques. A more robust feature selection procedure results from each phase that lessens the drawbacks of the one before it.

Keywords: Device communication patterns, Data breaches, IIoT device behavior, Tree-based techniques, Multiple perspectives assessment.

Introduction

IIoT systems have a big effect on safety, infrastructure, and industrial processes, so finding threats on them is very important. Often, IIoT systems manage important tasks in industries like healthcare, energy, transportation, and manufacturing. If these systems security is breached, it can cause terrible problems like damage to individuals and assets, as well as big financial losses. Good attack detection guarantees the integrity and dependability of these systems, therefore avoiding disturbances that can stop manufacturing processes or compromise private information. Early discovery ‹also enables quick reaction and mitigating action, therefore reducing the possible harm and recovery times. It also preserves the organization's reputation and helps to keep compliance with legal criteria. Maintaining operational continuity, guaranteeing safety, and preserving confidence in industrial processes depend on strong IIoT attack detection overall.

The proposed model tries to find the effective features in detection process of IIoT attacks. Effective feature selection is essential for accurately identifying threats in the IIoT using machine learning. The data generated by the IIoT generally consists of a substantial volume of data, some of which may be unrelated or repetitive when it comes to identifying potential attacks. Feature selection is a process that helps analyse this data by selecting the most useful elements that effectively differentiate between normal and harmful behavior. Consequently, there are several advantages to this approach, including the simplification of the ML model, resulting in improved speed and efficiency. Additionally, it aids in mitigating overfitting, a phenomenon in which the model excessively depends on certain intricacies within the training data, resulting in worse Performance on unknown data. In the end, carefully selected characteristics result in a more precise and dependable IIoT attack detection systems.

1.1 Data visualization techniques for feature selection

Data visualization approaches play a vital role in machine learning by aiding in the identification of the most relevant elements for a model via feature selection. Scatter plots may unveil associations and connections between features & the target variable. Heatmaps visually represent correlation matrices, showing features that have a significant association. This may help identify and remove duplicate features. Box plots display the spread & variability of features, aiding in the detection of anomalies. Pair plots provide a comprehensive display of scatter graphs for all combinations of features, unveiling the relationships and interactions between them. In addition, bar plots may display the distribution and significance of categorical variables, whereas histograms can illustrate the distribution and significance of continuous

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

features. By using these visual aids, individuals may make well-informed choices on which characteristics to keep, modify, or eliminate, so improving the overall performance and comprehensibility of the model.
Because it makes it easier to see the underlying relationships and structure in data, data visualization is essential for feature selection in machine learning. Through the use of visual aids, data scientists may discern patterns, connections, and anomalies in complicated datasets that are not readily discernible through statistical analysis in isolation. It is simpler to identify predictive factors when correlations between characteristics and the objective variable are revealed using methods like pair plots, heatmaps, and scatter plots. Furthermore, the pattern of distribution and deviation of features are displayed using visualizations like box plots and histograms, which help identify outliers and indicate what has to be changed or scaled down in order to improve model performance. Visual aids facilitate a thorough selection process by assisting in the comparison of feature relevance across different methods. Furthermore, data visualization helps ensure that the final model is reliable and effective by showing how modifications influence the accuracy and performance of the model. In the end, data visualization creates a bridge between the complexity of data and human comprehension, facilitating the creation of high-performing, transparent machine learning models and more precise, data-driven judgments. Different types of data visualization techniques are presented in figure-1

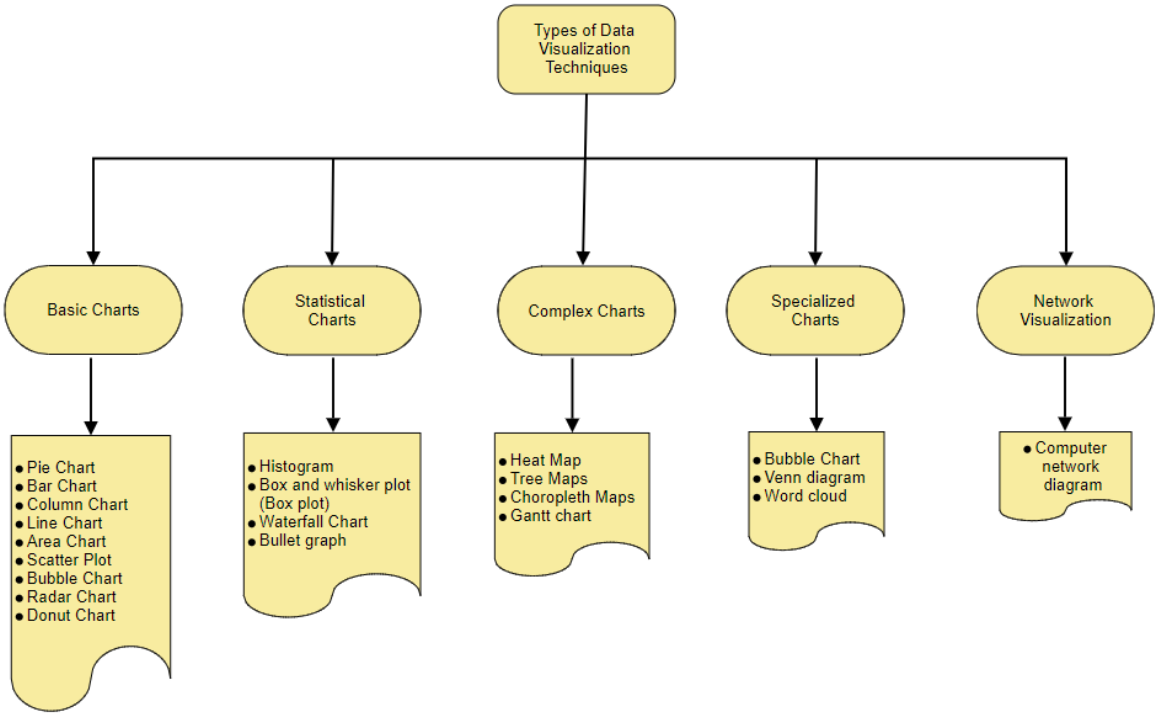


Figure 1: Types of different visualization techniques.

Data visualization is a useful tool for preliminary data exploration and comprehension, but it is not as good for accurate and scalable feature selection in large, complicated datasets due to its limits. To overcome these limitations and achieve more reliable feature selection, one way to address them is to integrate visualization with additional statistical and machine learning techniques. Complex non-linear correlations between characteristics and the goal variable may be difficult to see with visualization approaches.

1.2 Dynamic feature selection method using median and threshold values

The process of dynamic feature selection with median & threshold values entails the selection of features according to their statistical characteristics in relation to a predetermined threshold. Finding the median value for each attribute across all samples is the first step in this process. binding a threshold value which might be a set amount or a percentage of the median is the next stage. Then assess the features to check whether their median values are higher than this thresholds. Features whose medians are higher than the threshold are kept, while features whose medians are lower are eliminated. By continuously modifying the selection criteria in response to the features of the dataset, this method makes sure that the model contains only the most important elements, as judged by their overall tendency & relative relevance. By concentrating on the most important attributes, this strategy aids in decreasing dimensionality, strengthening model interpretability, and maybe improving performance.
There are benefits to employing median and threshold values in dynamic feature selection, particularly for particular kinds of datasets and situations. Consideration should be given to its drawbacks, which include threshold sensitivity, information loss, rigidity, disregard for interactions, computational complexity, reliance on high-quality data, and

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

restricted applicability. To overcome some of these limitations, this methodology can be combined with other feature selection techniques or used as part of a more comprehensive feature selection strategy.

2. Literature Survey

In [1], In this paper, feature importance is used to reduce the dimensions of IoT network traffic data, enabling lightweight stacking ensemble teaming to effectively identify network attacks with reduced storage space and minimal accuracy compromise. The growth of the Internet of Things industry is driven by the increasing number of connected devices. Combining single learners in stacking ensemble learning can improve classification performance. Feature importance combined with SEL enhances computational resource usage by focusing on important dataset features. The FI feature reduction technique identified the most important features for training the SEL classified. Different arrangements of single learners in the SEL classified have shown better performance in detecting IoT network intrusions. The SEL technique uses Decision Tree, Naive Bayes, and Logistic Regression to detect attacks efficiently. Using machine learning-based intrusion detection systems has improved classification accuracy in IoT networks. Collaborative intrusion detection models with ensemble classifiers have shown optimal results in reducing feature dimensions. Lightweight SEL classifiers have been effective in detecting normal and attack traffic in IoT datasets.

In [2], The study focused on developing a highly efficient intrusion detection system based on machine learning with a reduced set of 10 core features to effectively detect malicious attacks in industrial IoT applications, achieving high accuracy and recall rates exceeding 96% and 97% respectively. ln 'Feature Selection for Malicious Detection on Industrial IoT Using Machine Learning', Hao-J un Chen (2024) noted that the report by Microsoft highlighted the trend of integrating IoT and operational technology devices in organizations. A scheme combining PCC and frequency filtering rule was proposed to select core data features for machine learning models. LDA, RF, and CART were used for intrusion classification, and a hybrid KN N model was developed for intrusion detection. ML models were tested on 10 core features for detecting malicious attacks in an IIoT environment, achieving high accuracy. The 10 core features were effective in detecting and classifying attack types.

In [3], The Industrial Internet of Things (IIoT) faces privacy and security challenges, leading to the development of a self-attention-based deep convolutional neural network (SA-DCNN) model for monitoring and detecting malicious activities, which outperforms traditional machine learning and deep learning models. A research group from the Edinburgh Napier University led by Mohammed Alshehri (2024) studied a self-attention-based deep convolutional neural networks for IIoT networks intrusion detection. The text discusses the Industrial Internet of Things (IIoT) and proposes a self-attention- based deep convolutional neural network model for monitoring IIoT networks and detecting malicious activities. This model is evaluated using IoTID20 and Edge-IIoTset datasets, achieving high accuracy and outperforming traditional ML and DL models. Preprocessing steps, feature encoding, and performance comparisons with related articles are also discussed. The proposed method shows significant improvement in detecting cyberattacks in IIoT networks. The dataset contains 625,783 samples with 40,073 normal and 585,710 categorized into four types of attacks, further divided into eight sub-types, enhancing network intrusion detection capabilities.

In [4], Stakeholders expect the Industrial Internet of Things (IIoT) to be trustworthy and secure to prevent human casualties, leading to the need for novel security approaches like a SCADA- based cyberattack detection system use deep learning and decision tree algorithms. The text discusses a scalable and efficient DL- and decision tree-based ensemble cyberattack detection framework for SCADA-based IIoT networks. It emphasizes the importance of trustworthiness in the Industrial Internet of Things and proposes a reliable detection mechanism using deep and ensemble learning. The method was evaluated on 15 datasets and showed improved classification accuracy. The text also mentions the use of logistic regression, naive Bayes, random forests, and SVM in cyberattack detection. The proposed method aims to enhance trustworthiness by accurately detecting cyber attacks and selecting optimal features.

In [5], Data augmentation techniques are used to improve intrusion detection in Industrial Internet of Things (IIoT) networks, but the impact on classification performance varies depending on the algorithm and data used. In 'GPT and Interpolation-based Data Augmentation for Multiclass Intrusion Detection in IIoT', Francisco Melicias and colleagues (2024) noted that the text discusses the Industrial Internet of Things (IIoT) and the use of data augmentation techniques, such as Generative Pre-trained Transformers and interpolation, to generate IIoT network traffic. The evaluation showed mixed results in improving the performance of intrusion detection solutions, with XGBoost and Decision Trees classifiers performing the best. The text also mentions the use of Generative Adversarial Networks for data augmentation and the importance of evaluating data augmentation techniques before implementation. Additionally, the text highlights the challenges of class imbalance in data representation and the need for balanced datasets in IIoT traffic classification. The study involved 8 varied datasets. They advocate that further work includes evaluating the use of GANs for data augmentation in intrusion detection in IIoT and assessing its impact on classification algorithms with other IIoT network traffic datasets.

In [6], The integration of IT and OT networks in Industry 4.0 has increased vulnerabilities in Cyber—Physical Systems, leading to the development of a hybrid anomaly detection approach combining signature-based, threshold- based, and behavioral-based methods to improve accuracy and reduce detection time. Nicholas Jeffrey and colleagues (2024) reported in 'Using Ensemble Learning for Anomaly Detection in Cyber—Physical Systems' that society is transitioning

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

to Industry 4.0, Cyber- Physical Systems, which present unique challenges for threat detection. A hybrid methodology combining signature-based, threshold-based, and behavior-based detection strategies with Ensemble Learning is proposed for higher accuracy. The focus is on anomaly detection in Operational Technology (OT) networks within CPS environments. The use of Ensemble Learning shows minor accuracy improvements over traditional ML models, with the Bagging method showing slightly less ability to generalize across multiple CPS environments. The scarcity of training data for anomalous activity in CPS environments is a concern for operators due to potential financial and life safety impacts. The security measures for CPS on OT networks should target unforeseen behaviors related to physical components, with a focus on threshold- based detection strategies for immutable physical characteristics. The results suggest that Ensemble Learning can improve predictive performance in threat detection for CPSs compared to traditional ML methods. The results appear to differ from earlier work in this area: "Results in CPS environments can vary widely due to differences in features and data distribution, limiting the effectiveness of generic solution Ensemble Learning, which uses multiple ML algorithms on the same dataset, can improve predictive performance compared to using a single ML algorithm," Jeffrey argued.
In [7], The SecurityBERT architecture, utilizing the BERT model and privacy- preserving encoding techniques, outperforms traditional ML and DL methods in cyber threat detection, achieving 98.2% accuracy in identifying various attack types with minimal computational requirements and suitable for deployment on IoT devices. A research team led by Mohamed Ferrag of the University of Manchester (2024) reported on revolutionizing cyber threat detection with large language models. The global number of Internet of Things connected devices is projected to reach 30 billion by 2030. The Security BER|T model, utilizing a dataset, outperformed traditional Machine Learning models in identifying cyber threats with 98.2% accuracy. Researchers have explored using BERT in various cybersecurity applications. PPFLE and BBPE were used to represent network traffic data effectively. SecurityBERT was fine-tuned for cyber threat detection, achieving high accuracy. The model can be integrated into existing systems for improved detection rates. Future research can focus on enhancing SecurityBERT's performance and implementing autonomous mitigations based on its classifications. The team advocate that the paper suggests future research directions to enhance the use of LLMs in cybersecurity, including fine-tuning and expanding the SecurityBERT model to improve performance against different attack types and incorporating adversarial attacks. Continuous updating and training of the model on real-world datasets will be crucial to keep up with evolving cyber threats.

2.1. Research gaps identified:

1. Examine ways to improve resilience against zero-day assaults by continuously learning and adapting inside the ensemble framework.
2. Evaluate the latency and computational overhead of ensemble models in real-world deployments to ensure scalability and operational efficiency.
3. Investigate strategies to address the scarcity of training data for anomalous activities in CPS environments, which poses challenges for effective model training and generalization.

3. Proposed Methodology

Selecting features based on their significance, interpretability, and performance validation can be done in a complete way by integrating Random Forest, Integrated KNN, and Logistic Regression. When developing accurate and dependable predictive models across a range of domains, a multi-model approach can result in more robust and efficient feature selection. Every approach verifies and improves the feature selection procedure, lowering the possibility of overfitting and guaranteeing that the features chosen are pertinent to various modeling modalities. The process of proposed methodology is shown in figure-2

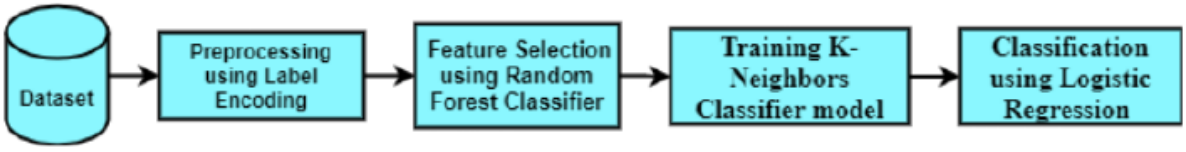


Figure 2: Proposed Methodology for Feature Selection

3.1. Label Encoding: Label encoding is a step before machine learning that turns category data into numbers so that programs can use them. To put it simply, it gives each group in a feature a unique number. This is very helpful for algorithms that need numerical input, like many neural network models and regression techniques. One of the best things about label encoding is how easy and quick it is to work with categorical data. There may not be an underlying numeric link between the groups because of the limitation. Label encoding has some problems, but it can be very useful in some situations where the category data is automatically ordered. When working with non-ordinal category textures, it's best to avoid adding fake ordinal links by using methods like one-hot encoding. The right encoding is very important

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

for model success, since the wrong encoding can cause models to be inaccurate and give false results. The process is shown in figure 3.

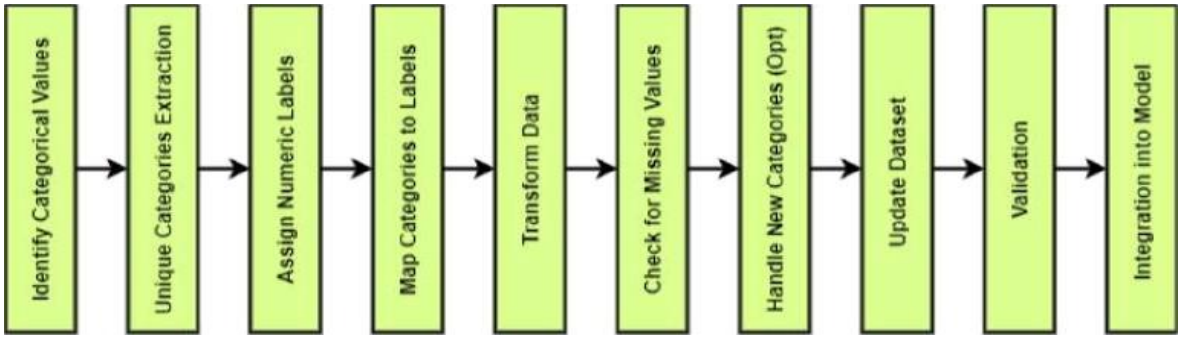


Figure 3: Label Encoding Process

3.2. Random Forest Classifier for feature selection:

Feature selection is facilitated by the Random Forest Classifier, which is capable of ranking features in order of importance. During training, it builds many decision trees and averages each one's forecast. Based on how successfully a feature decreases impurity (such as entropy or Gini impurity) at each split, the split choices made by each tree provide information about the value of a feature. crucial features are those that are routinely employed by many trees to make crucial judgements. Metrics such as the mean drop in accuracy or impurity when an attribute is removed may be used to quantify its significance. Through the Random Forest Classifier, features me ranked according to these criteria, which helps determine which characteristics have the most influence and enhances the performance and efficiency of the model. The computation or feature importance is shown in equation (1)

$$\mathrm{RFI_k} = \frac{1}{\mathrm{n}} * \sum_{\mathrm{k=1}}^{\mathrm{m}} \mathrm{RFimpurity(t, F_k)} * \frac{\mathrm{n_t}}{\mathrm{N}} - (1)$$

n represents number of trees
t isdecrease function at node
F_k splitting kth feature at node t
n_t is number of samples at node t
N is total number of samples

Because Random Forest naturally measures feature relevance during the construction of predictive models, it is an effective ensemble learning technique for feature selection. During training, Random Forest builds a large number of decision trees and aggregates the output to improve prediction accuracy and reveal the relative significance of each attribute. Based on how much each feature lowers impurity (such as Gini impurity or entropy) across all trees in the forest, the method determines feature significance scores. Features that have a major impact on lowering impurity are valued higher. This procedure aids in locating and prioritizing the most significant elements in a dataset and is frequently represented visually using significance plots. Moreover, Random Forest is especially well-suited for feature selection due to its resistance to overfitting and capacity to manage big datasets with greater dimensionality. Data scientists may streamline their models, lower processing costs, and even enhance model generalization by concentrating on the highest-ranked characteristics. Furthermore, Random Forest can manage varied interactions, capturing intricate linkages that more straight forward approaches could overlook. Random Forest is a useful technique in the machine learning feature selection process since this thorough evaluation guarantees that the chosen features are not only pertinent but also enhance the prediction ability of the model. The entire process is shown in figure 4.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

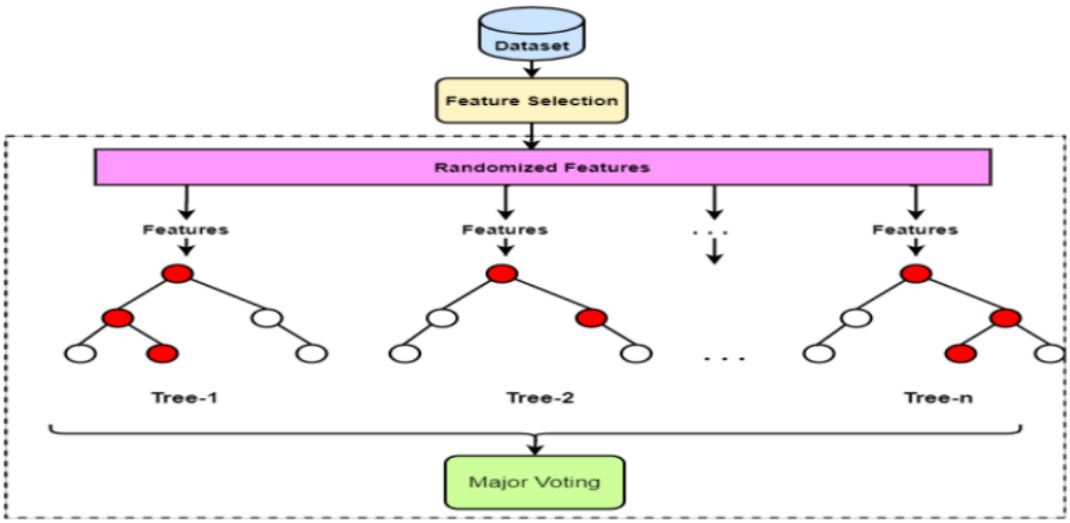


Figure 4: Feature Selection using Random Forest

3.3. Training K-Neighbors Classifier model for selection of features

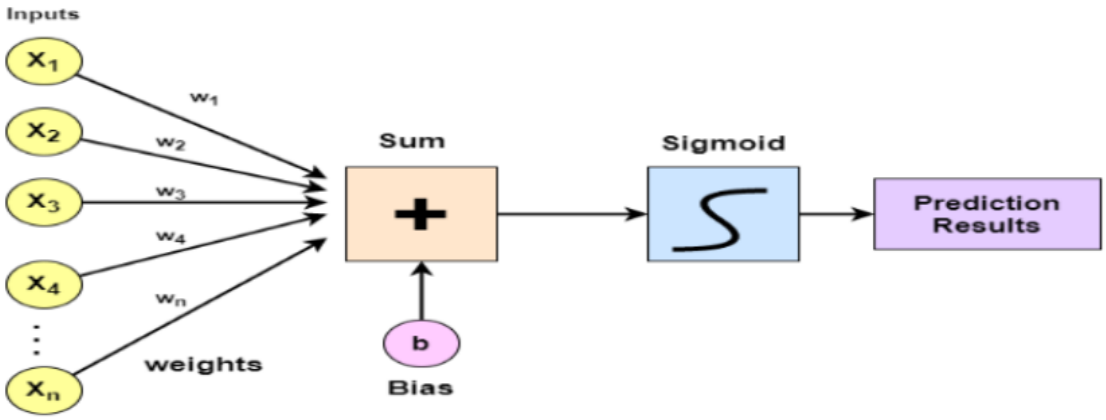
By calculating the separation between data points in the feature space, KNN determines the significance of a feature. Data point proximity is often impacted by features that play a major role in class distinction. Features that have a greater impact on the distance computations are deemed more significant. To learn more about the structure of the dataset, investigate the significance of various features with respect to the classification problem. Using data with certain traits to train a KNeighbors Classification model takes more than one step. Use feature selection to choose the most significant characteristics from the dataset. Next, divide the information into sets for training and testing to check how well the model works. Set up the KNeighbors Classifier by giving il the amount of neighbors (k) and any other hyperparameters that it needs. Training the model with the training set lets it learn patterns from specified features. Check the model's validation measures on the test data set after training it. To enhance performance, change the features or hyperparameters as needed. Finally, make predictions using the trained model on new data, making sure to use tire same features you chose to be consistent. This process makes sure that the model works well and efficiently, using the best data for the best results.

3.4. Logistic Regression for feature selection:

Logistic regression is a useful tool for feature selection in machine learning pipelines, especially when simplicity and interpretability are important considerations. The strength of the correlation between each characteristic and the target variable (log-odds) in logistic regression is shown by the size of the coefficients. Larger coefficients (in absolute terms) indicate that a feature is more significant for outcome prediction.

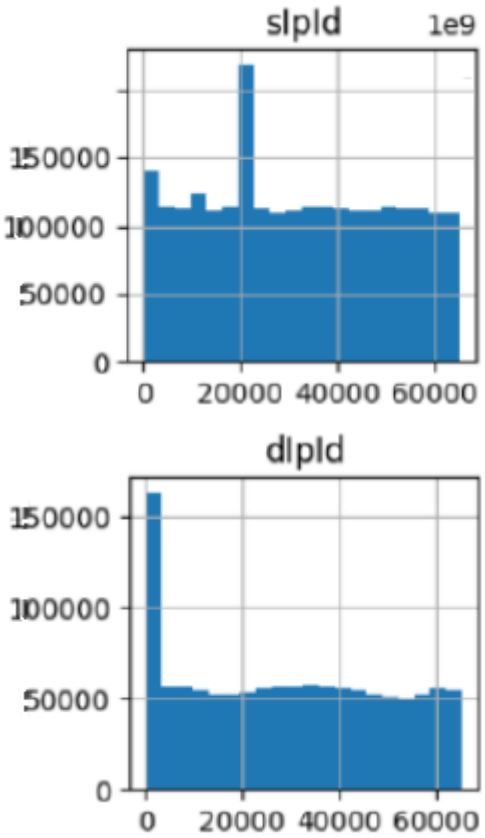
Logistic regression is a key machine learning classification approach for categorical predictions. Through the use of a logistic function, it creates a model that represents the chance that a certain input conforms to a classification. Linear decision boundaries are estimated by the method using input characteristics and result log-odds. predicts the input's class in binary classification. In the case that this likelihood is higher than a certain threshold, the model will allocate the input to the category in question; otherwise, it will assign it to the other category. Maximum likelihood estimation estimates model parameters for the beer training data fit. logistic Regression is efficient and interpretable, making it versatile. It supports multiclass classification using OvR or softmax regression. Regularisation met hods like L1 lasso & L2 ridge avoid overfitting and enhance generalisation.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

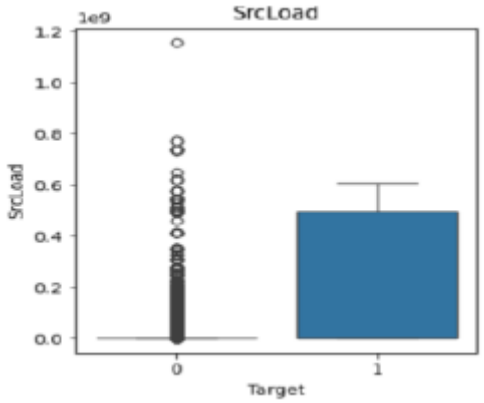


4. Results & Discussion

Figure 5 presents the data visualization techniques of the different features along with the statistics of different numerical features. These help us to correlate the relation between the features along with the impact of each feature on the prediction variable.



REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024



Traffic	
normal	1107448
DoS	78305
Reconn	8240
CommInj	259
Backdoor	212
Name:couns,	dtype:int64

Figure 5: Data Visualization of features

Table 1 presents the important features of IIoT attacks by computing their score using the random forest algorithm. The significance of a feature in Random Forests is determined by its contribution to the reduction of impurity or error in the decision trees inside the forest. Larger values, or features with higher importances, are thought to have stronger predictive power for the target variable.

Table 1: Feature Importance by Random Forest

Number	**Feature**	**Importance**
43	Traffic	0.181091
22	DIntPkt	0.140365
36	DAppBytes	0.094370
4	DstPkts	0.088692
6	DstBytes	0.085351
20	DstJitter	0.067587
38	SynAck	0.062398
25	TcpRtt	0.052429
31	sTtl	0.037341
13	DstRate	0.035931
16	DstLoss	0.028210
34	dIpId	0.026729
32	dTtl	0.017959
1	Sport	0.015657
10	DstLoad	0.013811
17	Loss	0.013786

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

Number	Feature	Importance
2	Dport	0.007060
37	TotAppByte	0.006927
18	pLoss	0.006627
9	SrcLoad	0.002343
26	IdleTime	0.002215
33	sIpId	0.001927
23	Proto	0.001880
7	SrcBytes	0.001709
35	SAppBytes	0.001427
11	Load	0.001418
8	TotBytes	0.000842
5	TotPkts	0.000839
24	Dur	0.000737
15	SrcLoss	0.000535
3	SrcPkts	0.000521
12	SrcRate	0.000362
28	Min	0.000203
42	DstJitAct	0.000174
14	Rate	0.000173
21	SIntPkt	0.000085
27	Sum	0.000076
29	Max	0.000062
41	SrcJitAct	0.000042
39	RunTime	0.000036
19	SrcJitter	0.000032
30	sDSb	0.000024
40	sTos	0.000013
0	Mean	0.000003

Figure 6 presents the integrated approach in which it determines the every characteristic that contributes equally to the distance computations by normalizing the data. Using the dataset as training data create a KNN model and assign a score to each feature according to how relevant it is to the classification problem. KNN provides metrics for feature significance and distance that can be used for this purpose. To determine which features are highest ranked, use the KNN scores or rankings as a guide. Evaluate the coefficients allocated to each feature and train a Logistic Regression model using the chosen features. When regularization is applied, features with non-zero coefficients are deemed to be

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 01 (2026)
http://www.veterinaria.org
Received 01 June 2023 Revised 19 March 2024 Accepted 29 March 2024

```
Index(['Sport', 'Dport', 'SrcPkts', 'DstPkts', 'TotPkts', 'DstBytes',
       'SrcBytes', 'TotBytes', 'SrcLoad', 'DstLoad', 'Load', 'SrcRate',
       'DstRate', 'SrcLoss', 'DstLoss', 'Loss', 'pLoss', 'DstJitter',
       'DIntPkt', 'Proto', 'Dur', 'TcpRtt', 'IdleTime', 'Min', 'sTtl', 'dTtl',
       'sIpId', 'dIpId', 'SAppBytes', 'DAppBytes', 'TotAppByte', 'SynAck',
       'Traffic'],
      dtype='object')
```

chosen.

Figure 6: features generated by Proposed Multi Model

Figure 7 determines performance metrics of the proposed model based on the fo1low'ing equation.

$$Multi\_Model_{Acc} = \frac{TP_{Attack}+TN\_Attack}{TP_{Attack}+FP_{Nonattack}+FN_{Nonattack}+TNAttacl} \text{ - (2)}$$

Metric	Logistic Regression	k-Nearest Neighbors
Accuracy	0.9306	0.9979
Precision	0.9301	0.9978
Recall	0.9103	0.9974
F1-score	0.9201	0.9976

Figure 7: Metrics evaluation using the proposed Methodology

Conclusion

IIoT systems, which handle sensitive data and operational procedures, are essential to many different sectors. In order to avoid disruptions, monetary losses, and data breaches, it is essential to protect these systems from cyberattacks. The efficacy of the model may suffer if conventional feature selection techniques like LASSO ignore correlated features. In order to refine feature selection based on various modeling perspectives—ensemble, distance-based, and linear this work suggests a robust method that uses Random Forest for initial feature ranking and KNN and Logistic Regression thereafter. By balancing efficiency, interpretability, and complexity, our multi-model strategy improves IIoT attack detection accuracy while reducing overfitting. Data visualization promotes transparency and efficiency in models by helping to find pertinent features. Future studies ought to concentrate on strengthening defenses against changing threats and maximizing computational effectiveness for practical use.

References

1. Abdulkareem, Sulyman and Foh, Chuan and Carrez, François and Moessner, Klaus, A Lightweight Sel for Attack Detection in Iot/Iiot Networks. Available at SSRN: https://ssrn.com/abstract=4772831 or http://dx.doi.org/10.2139/ssrn.4772831
2. Chuang, Hong-Yu & Chen, Ruey-Maw. (2024). Feature Selection for Malicious Detection on Industrial IoT Using Machine Learning. Sensors and Materials. 36. 1035. 10.18494/SAM4666.
3. Alshehri, Mohammed & Saidani, Oumaima&Alrayes, Fatma & Abbasi, Saadullah & Ahmad, Jawad. (2024). A Self-Attention-Based Deep Convolutional Neural Networks for IIoT Networks Intrusion Detection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3380816.
4. Supriya, Belkar. "Trustworthy and reliable machine learning based cyberattack detection in IOT." IJPAST 14, no.1 (2024)
5. Melicias, Francisco & Ribeiro, Tiago &Rabadão, Carlos & Santos, Leonel & Costa, Rogério Luís. (2024). GPT and Interpolation-Based Data Augmentation for Multiclass Intrusion Detection in IIoT. IEEE Access. 12. 17945-17965. 10.1109/ACCESS.2024.3360879.
6. Jeffrey, Nicholas & Tan, Qing & Villar, Jose. (2024). Using Ensemble Learning for Anomaly Detection in Cyber–Physical Systems. Electronics. 13. 1391. 10.3390/electronics13071391.
7. Ferrag, Mohamed Amine & Ndhlovu, Mthandazo& Tihanyi, Norbert & Cordeiro, Lucas & Debbah, merouane&Lestable, Thierry & Thandi, Narinderjit. (2024). Revolutionizing Cyber Threat Detection with Large Language Models: A privacy-preserving BERT-based Lightweight Model for IoT/IIoT Devices. IEEE Access. PP. 10.1109/ACCESS.2024.3363469.
8. Zhang, Y., et al. (2023). A survey on machine learning for anomaly detection in IIOT networks. Computers &Eletrical Engineering. https://doi.org/10.1013/j.compeleceng.2023.107670

9. Jin, X.,et al.(2023). Deep learning-based intrusion detection system for IIoT networks. Further Generation Computer Systems. https://doi.org/10.1013/j,future.2023.08.022
10. Lee, S., &Kim,J.(2022 Ensemble-based intrusioin detection system for industrial IoT environments. Computers & Security, 114, 102293. https//doi.org/10.1016/j.cose.2021.102293