Science and Technology Publishing (SCI & TECH)
ISSN: 2632-1017
Vol. 10 Issue 2, February - 2026

Applied Extreme Gradient Boosting Model for anomaly detection using different feature subsets on Industrial Internet of Things network dataset

Michael Oghale Ighofiomoni

Department of Computer Engineering
Southern Delta University, Ozoro, Delta State, Nigeria
Ighofiomonimo@dsust.edu.ng

Abstract

Applied Extreme Gradient Boosting (XGBoost) model for anomaly detection using different feature subsets on Industrial Internet of Things network dataset is presented. The Industrial Internet of Things Dataset (IIoTID) used had 79 features. The study examined the impact of feature subset on the classification performance and execution time of XGBoost model employed in anomaly detection on IIoT networks. The XGBoost was trained and validated using different feature subsets in the dataset selected using the feature importance ranking and the model accuracy and execution time sensitivity analysis. The results showed that the highest accuracy of 99.85 % occurred when 60 or more features are used. At the same time, the model training time also is at its highest value of 53 seconds when the 60 or more features are used. Also, at exactly 20 features subset, the XGBoost model accuracy was 99.8 % with training time of 2.76 seconds. On the other hand, at the 60th feature, the model accuracy was 99.88 % with training time of 6.53 seconds. This shows that the system resources can be preserved by adopting 20 feature subset without losing any significant value in the accuracy of the model. In all, the feature subset consisting of the top 20 to 22 features was recommended as the most appropriate for the case study XGBoost model and IIoTID dataset.

Keyword: Extreme Gradient Boosting Model, Anomaly Detection, Threat Classification, Feature Subsets Selection, Industrial Internet of Things (IIoT) network, Feature Importance Ranking

1. Introduction

Nowadays, Artificial Intelligence (AI) models are used in diverse field to facilitate automation and build smart autonomous systems [1,2 3]. Such applications are seen in autonomous vehicles, smart grids, smart agriculture and smart manufacturing systems [4,5]. Such systems depends heavily on the use of sensors, microcontroller and communication links that connect them to the Internet [6,7]. Such a network is termed Internet of Things, when designed specifically for the industrial applications, they are referred to as Industrial Internet of Things (IIoT) [8,9].

Today, the IIoT has gained wide applications resulting in autonomous industrial process and manufacturing where only robots are used without human involvement [10,11]. While those IIoT networks are very crucial to the automation systems, they are very susceptible to cyber-attacks [12,13]. The sensors and many other components of the IIoT networks are resource constrained, thereby limiting

the range of security features that can be implemented on them [14,15]. Hence, the IIoT security is a big issues in the modern industrial applications. Any solution developed should consider the resource constraint prevalent in IIoT [17,18].

Accordingly, in this present study, the use of XGBoost for anomalous traffic detection is studied. Specifically, given that the IIoT intrusion detection dataset is always loaded with too many features set, the focus in this work is to evaluate the XGBoost performance using different carefully selected feature subset and thereby identify the best feature subset that will give the optimal threat identification accuracy with minimal execution time [19,20]. This study is therefore specifically useful for such resource constrained networks like the IIoT networks.

2. Methodology

2.1 The case study dataset and feature set

The study examined the impact of feature subset on the classification performance and execution time of Extreme Gradient Boosting (XGBoost) model employed in anomaly detection on IIoT networks. The XGBoost is trained and validated using different feature subsets in the dataset.

Specifically, the Industrial Internet of Things Dataset (IIoTID) is used in the study and after data cleaning and some preprocessing, about 79 features are identified, as listed in Table 1. Feature importance ranking was conducted using the XGBoost model. The features ranked based on their importance scores are then selected in steps of 10, starting with the top 10 ranked features, top 20 ranked features, and son till all the 79 features are considered. The model prediction performance is recorded in each training and validation instance along with the model training time. The model performance for the eight different feature subset selections are then compared to discuss the effect of the feature selection on the XGBoost classification of the anomalous data classes in the IIoTID dataset.

Figure 1 The 79 features are identified the Industrial Internet of Things Dataset (IIoTID)

S/N	Feature	S/N	Feature	S/N	Feature
1	Scr_bytes_ratio	28	is_syn_only	55	Avg_iowait_time
2	Des_bytes_ratio	29	Avg_ideal_time	56	Avg_wtps
3	Des_port	30	Protocol_udp	57	OSSEC_alert_level
4	byte_rate	31	is_pure_ack	58	Std_ideal_time
5	Scr_bytes	32	Service_dns	59	Std_iowait_time
6	Avg_num_cswch/s	33	Std_num_proc/s	60	Service_ssh
7	paket_rate	34	is_with_payload	61	Avg_tps
8	Scr_ip_bytes	35	Is_SYN_ACK	62	Std_kbmemused
9	Service_other	36	Avg_nice_time	63	Avg_rtps
10	Duration	37	Std_user_time	64	anomaly_alert_True
11	Des_bytes	38	Conn_state	65	Std_rtps
12	total_bytes	39	anomaly_alert_TRUE	66	Std_wtps
13	Des_pkts_ratio	40	Scr_port	67	anomaly_alert_FALSE
14	Service_mqtt	41	Avg_ldavg_1	68	Std_tps
15	Des_ip_bytes	42	Login_attempt	69	Service_smtp
16	Avg_user_time	43	File_activity	70	Service_private
17	total_packet	44	Process_activity	71	Service_simple_service_discovery
18	Des_pkts	45	Service_websocket	72	Service_netbios-ns
19	std_num_cswch/s	46	Avg_kbmemused	73	Service_echo
20	Avg_num_Proc/s	47	Service_modbus	74	Service_imap
21	read_write_physical.process	48	Std_nice_time	75	missed_bytes
22	Scr_packts_ratio	49	Std_system_time	76	Service_dhcp
23	Scr_pkts	50	OSSEC_alert	77	Service_mysql
24	Service_http	51	Service_https	78	is_SYN_with_RST
25	FIN or RST	52	Succesful_login	79	Bad_checksum
26	Avg_system_time	53	is_privileged		
27	Protocol_tcp	54	Std_ldavg_1		

2.2 The XGBoost model,

The XGBoost model, with the architecture in Figure 1, is an optimized distributed gradient boosting library designed for high performance and flexibility. It implements machine learning algorithms under the Gradient Boosting framework and is known for its speed and accuracy.

The XGBoost builds trees one after another, where each new tree tries to fix the mistakes of the previous ones. This boosting process makes the model especially effective at handling complex patterns in the data. In this research, all 79features are considered. XGBoost chooses the ones that most improve accuracy by reducing the model's current error. The predictions are generated by

adding up the outputs of all the trees in the model. For classification, this sum is passed through a function (like sigmoid) to produce probabilities, which are then converted to class labels. The key parameter settings used for the DTM are as follows;

(i) `random_state=42`

(ii) `objective='binary:logistic'` → suited for binary classification tasks

(iii) `n_estimators=100` → number of trees added sequentially

(iv) `learning_rate=0.1` → controls the contribution of each new tree

(v) `max_depth=6` → limits how deep each tree can go

(vi) `subsample=1.0`, `colsample_bytree=1.0` → use all data and features by default

(vii) `gamma=0` → no regularization for tree splits by default

The XGBoost model was trained using 70% of the dataset while the validation was done using 30 of the dataset.

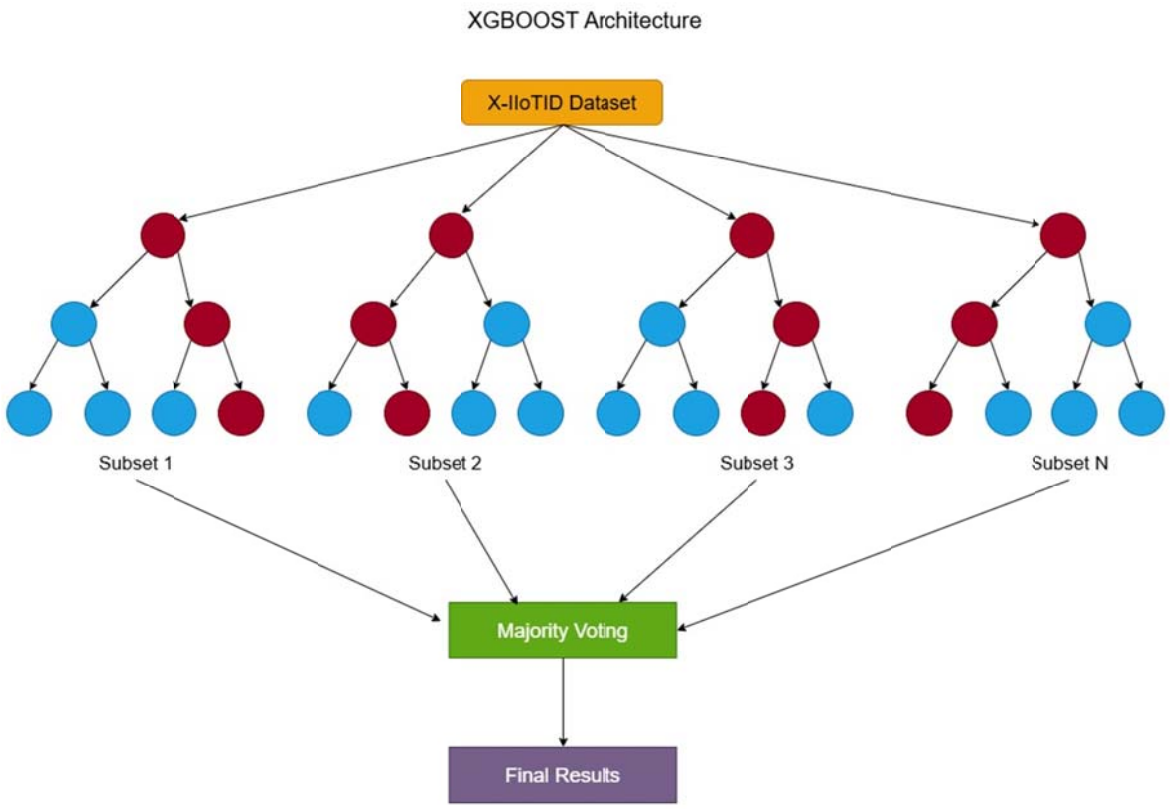

Figure 1 The Architecture of XGBoost model

3. Results and discussion

3.1 The Feature Ranking Results of the XGBoost Model

In view of the very large number of features in the dataset, the results of the feature ranking are segmented into 4 sections, namely

(i) The features with ranking score above 0.002 or above 0.21 % score (Table 2 and Figure 2) . This group has the first 20 features with Des _port as the number 1 ranked feature with a score of 0.162934. The least feature in this group is ranked 20 with a score of 0.002095.92

(ii) The features with ranking score above 0.00015 or 0.015 % score but below 0.002 or 0.21 % (Table 3 and Figure 3). This group has the 21st to 37th features with Des _bytes as the number 21 ranked feature with a score of 0.00163223. The least feature in this group is ranked 37 with a score of 0.000157024

(iii) The features with ranking score above 0.00001` or 0.01 % score but below 0.00015 or 0.015 % (Table 4 and Figure 4). This group has the 28th to 54th features with Service dns as the number 28^{th} ranked feature with a score of 0.000139746. The least feature in this group is ranked 54 with a score of 0.0000517157

(iv) The features with ranking score below 0.00001` or 0.01 % (Table 5 and Figure 5). This group has the 59th to 79th features with Avg_Ideal_time as the number 59^{th} ranked feature with a score of 0.000049374. The least feature in this group is ranked 79 with a score of 0.0

Table 2 The list of the features with ranking score above 0.002 or above 0.21 % score

Feature Rank	Feature S/N	Feature	Feature Importance Rank Score (in fraction) Based on XGBoost	Feature Importance Rank Score (in %) Based on XGBoost
1	1	Des_port	0.162934	16.29%
2	3	Scr_bytes	0.158293	15.83%
3	48	Avg_num_cswch/s	0.142719	14.27%
4	21	byte_rate	0.10739	10.74%
5	46	Avg_num_Proc/s	0.0948063	9.48%
6	20	paket_rate	0.0835176	8.35%
7	58	Protocol_tcp	0.0800167	8.00%
8	70	Service_other	0.0381545	3.82%
9	67	Service_mqtt	0.0271212	2.71%
10	66	Service_modbus	0.0247267	2.47%
11	56	read_write_physical.process	0.0178747	1.79%
12	0	Scr_port	0.0119952	1.20%
13	24	Scr_bytes_ratio	0.00840789	0.84%
14	17	Des_ip_bytes	0.00738359	0.74%
15	28	Avg_nice_time	0.00545758	0.55%
16	19	total_packet	0.00369378	0.37%
17	18	total_bytes	0.00365475	0.37%
18	73	Service_smtp	0.00339686	0.34%
19	42	Avg_ldavg_1	0.00290717	0.29%
20	2	Duration	0.00209592	0.21%

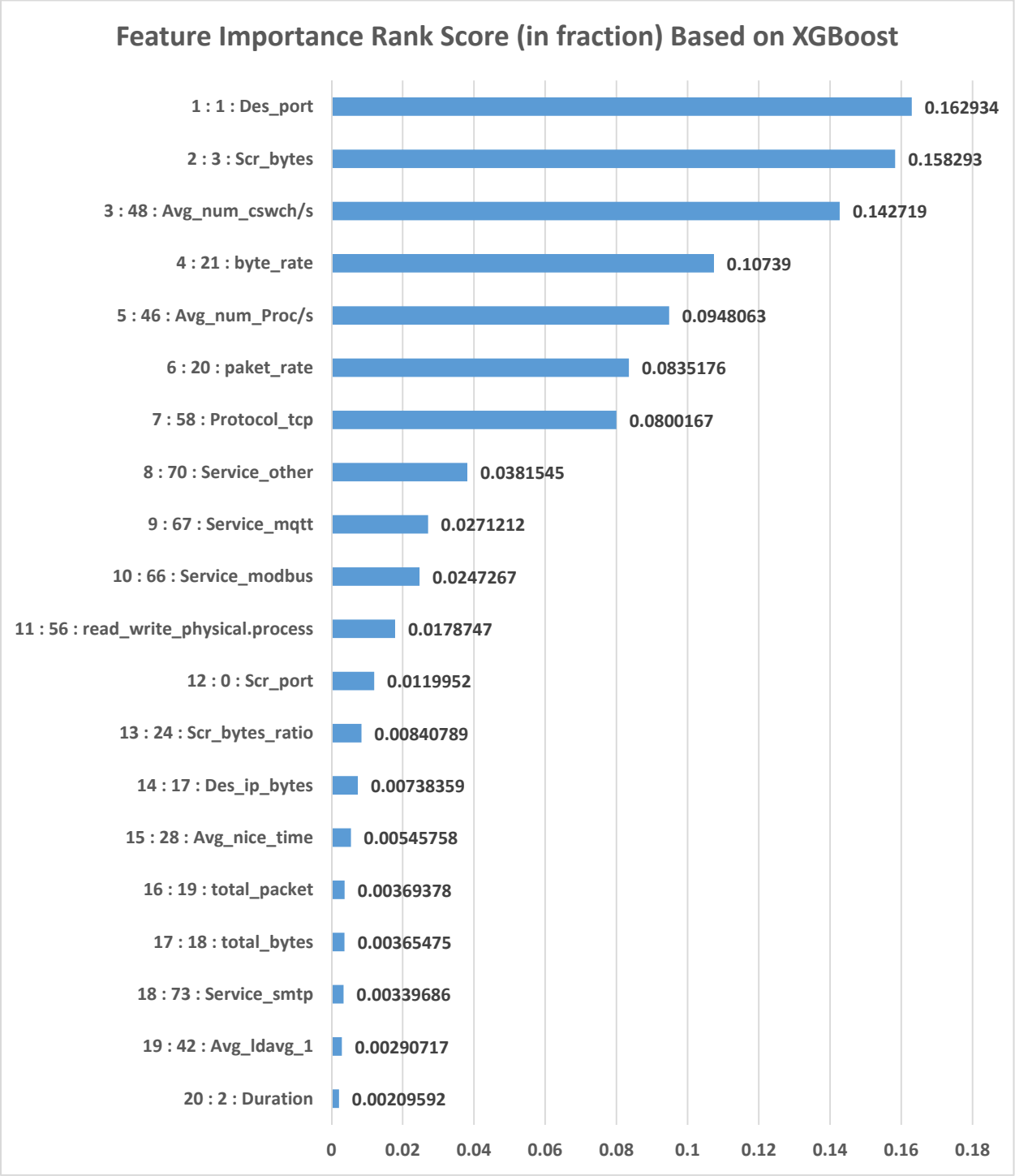

Figure 2 Bar chart of the features with ranking score above 0.002 or above 0.21 % score

Table 3 The list of the features with ranking score above 0.00015 or 0.015 % score but below 0.002 or 0.21 %

Feature Rank	Feature S/N	Feature	Feature Importance Rank Score Based on XGBoost	Feature Importance Rank Score Based on XGBoost
21	4	Des_bytes	0.00163223	0.16%
22	14	Scr_pkts	0.00158295	0.16%
23	27	Std_user_time	0.00153779	0.15%
24	32	Avg_iowait_time	0.0012135	0.12%
25	47	Std_num_proc/s	0.000792759	0.08%
26	15	Scr_ip_bytes	0.000789919	0.08%
27	22	Scr_packts_ratio	0.000787131	0.08%
28	59	Protocol_udp	0.000596618	0.06%
29	25	Des_bytes_ratio	0.000555985	0.06%
30	57	is_privileged	0.000525198	0.05%
31	9	is_pure_ack	0.000306665	0.03%
32	16	Des_pkts	0.000299067	0.03%
33	51	OSSEC_alert_level	0.000258967	0.03%
34	26	Avg_user_time	0.000201653	0.02%
35	35	Std_ideal_time	0.000174107	0.02%
36	44	Avg_kbmemused	0.000161112	0.02%
37	49	std_num_cswch/s	0.000157024	0.02%

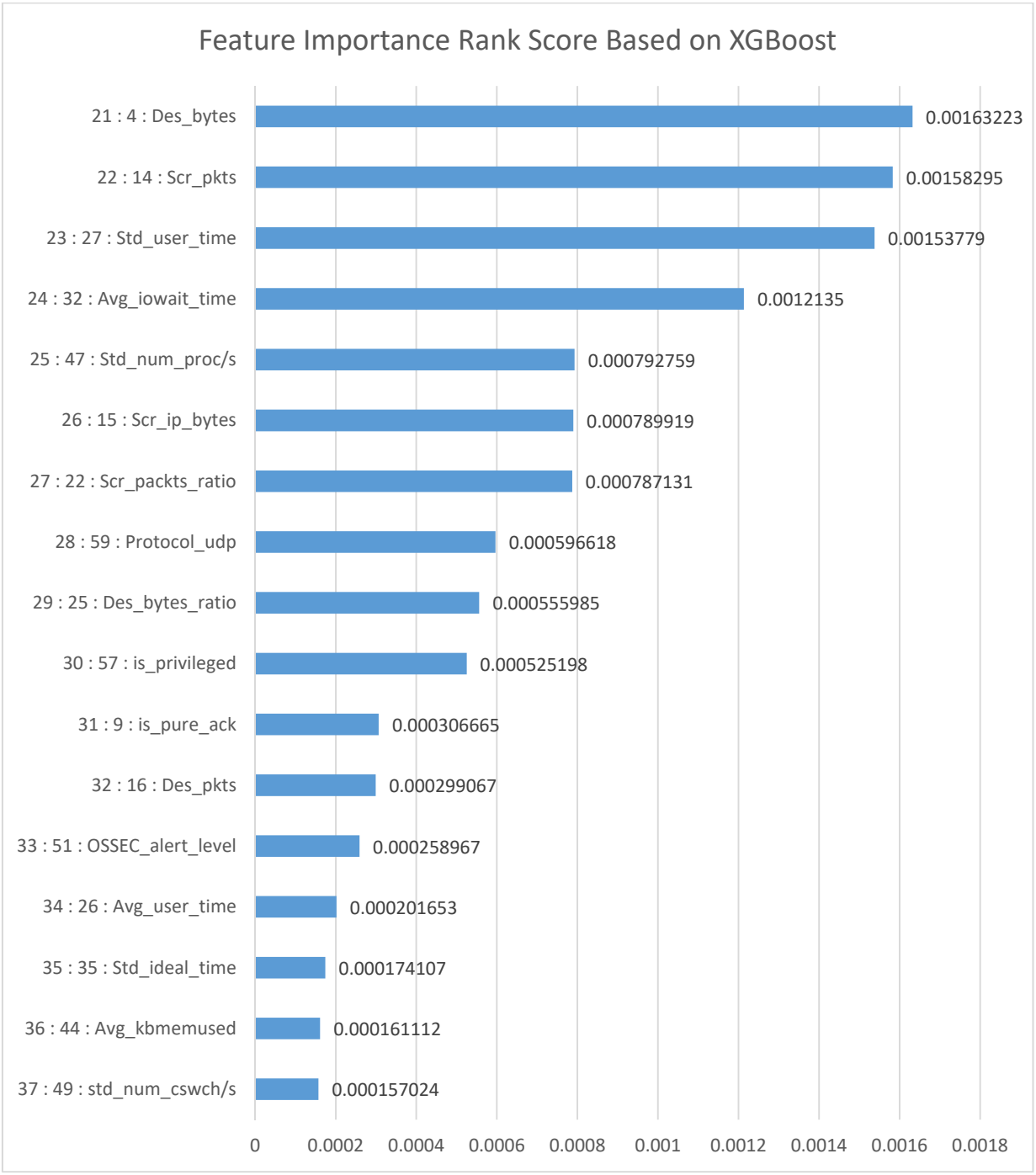

Figure 3 Bar chart of the features with ranking score above 0.00015 or 0.015 % score but below 0.002 or 0.21 %

Table 4 The list of the features with ranking score above 0.00001` or 0.01 % score but below 0.00015 or 0.015 %

Feature Rank	Feature S/N	Feature	Feature Importance Rank Score Based on XGBoost	Feature Importance Rank Score Based on XGBoost
38	61	Service_dns	0.000139746	0.01%
39	10	is_with_payload	0.000131223	0.01%
40	75	Service_websocket	0.000121141	0.01%
41	5	Conn_state	0.000120394	0.01%
42	63	Service_http	0.000115652	0.01%
43	30	Avg_system_time	0.000109392	0.01%
44	8	Is_SYN_ACK	0.000106975	0.01%
45	7	is_syn_only	0.000105722	0.01%
46	43	Std_ldavg_1	8.86256E-05	0.01%
47	45	Std_kbmemused	7.98484E-05	0.01%
48	29	Std_nice_time	7.73919E-05	0.01%
49	31	Std_system_time	7.44672E-05	0.01%
50	41	Std_wtps	6.81348E-05	0.01%
51	23	Des_pkts_ratio	6.50485E-05	0.01%
52	55	Process_activity	5.89926E-05	0.01%
53	40	Avg_wtps	5.80904E-05	0.01%
54	39	Std_rtps	5.17157E-05	0.01%

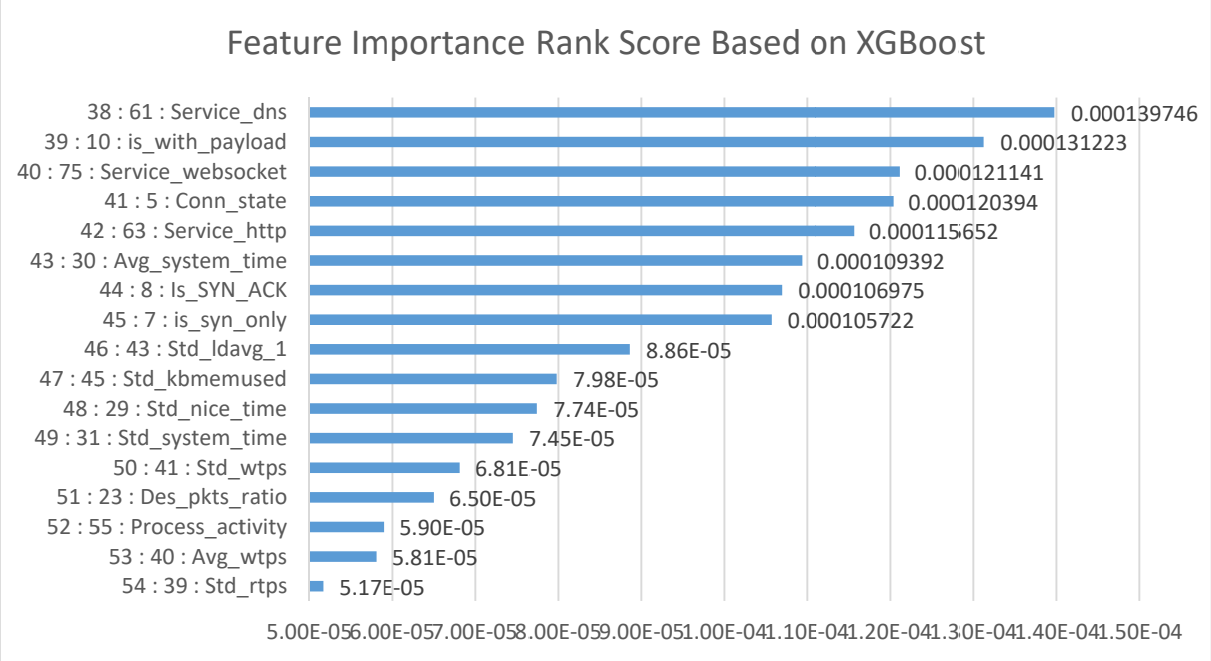

Figure 4 Bar chart of the features with ranking score above 0.00001` or 0.01 % score but below 0.00015 or 0.015 %

Table 5 The list of the features with ranking score below 0.00001` or 0.01 %

Feature Rank	Feature S/N	Feature	Feature Importance Rank Score Based on XGBoost	Feature Importance Rank Score Based on XGBoost
55	34	Avg_ideal_time	4.93784E-05	0.00%
56	38	Avg_rtps	4.39854E-05	0.00%
57	33	Std_iowait_time	3.59985E-05	0.00%
58	37	Std_tps	3.57647E-05	0.00%
59	36	Avg_tps	3.53889E-05	0.00%
60	74	Service_ssh	2.66013E-05	0.00%
61	54	File_activity	2.10911E-05	0.00%
62	50	OSSEC_alert	1.91126E-05	0.00%
63	78	anomaly_alert_TRUE	1.90713E-05	0.00%
64	11	FIN or RST	0.000007683	0.00%
65	77	anomaly_alert_FALSE	5.4487E-06	0.00%
66	72	Service_simple_service_discovery	0.000003202	0.00%
67	69	Service_netbios-ns	2.1347E-06	0.00%
68	52	Login_attempt	2.0355E-06	0.00%
69	76	anomaly_alert_True	7.116E-07	0.00%
70	6	missed_bytes	0	0.00%
71	12	Bad_checksum	0	0.00%
72	13	is_SYN_with_RST	0	0.00%
73	62	Service_echo	0	0.00%
74	53	Succesful_login	0	0.00%
75	60	Service_dhcp	0	0.00%
76	68	Service_mysql	0	0.00%
77	65	Service_imap	0	0.00%
78	64	Service_https	0	0.00%
79	71	Service_private	0	0.00%

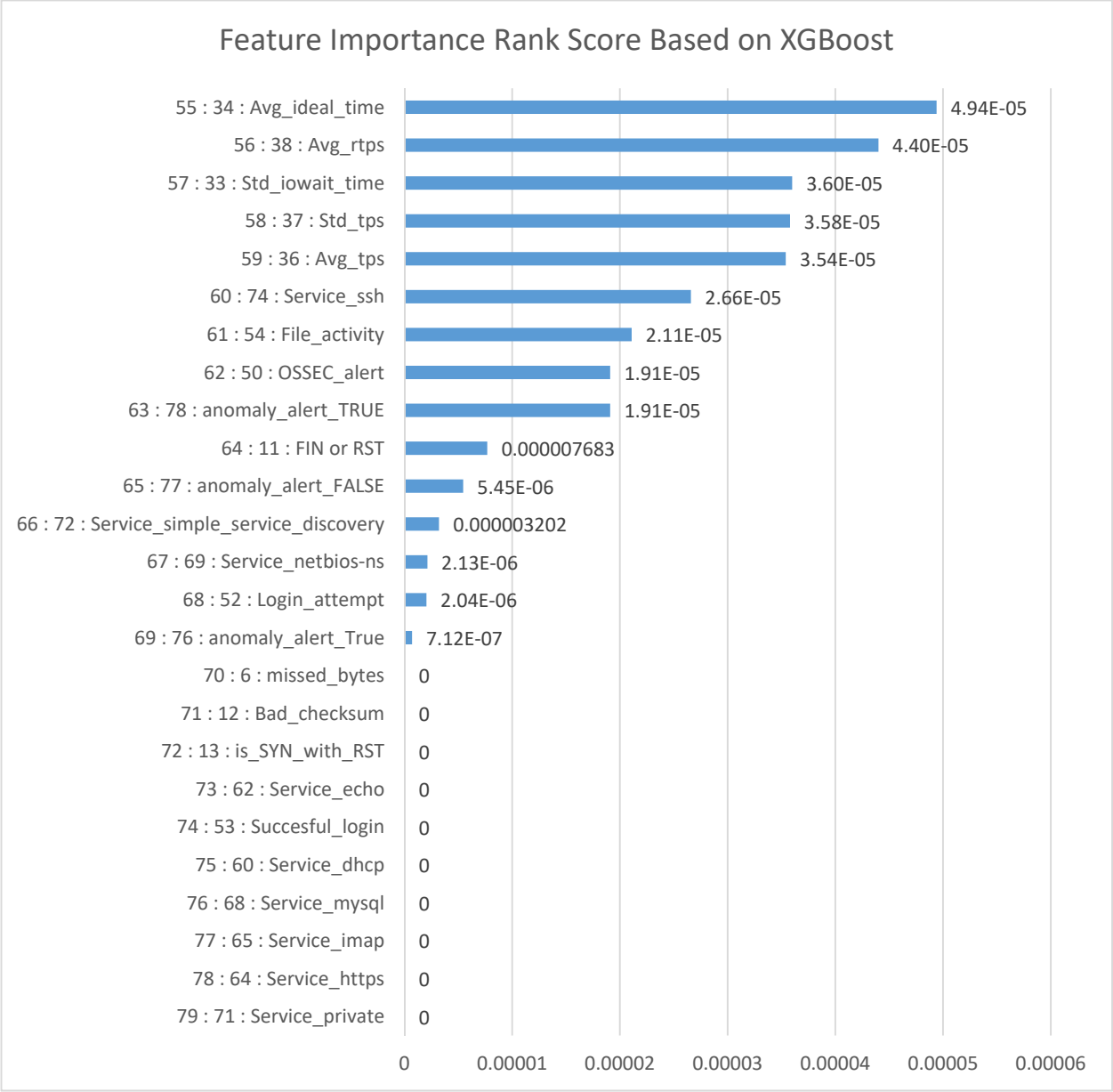

Figure 5 Bar chart of the features with ranking score below 0.00001` or 0.01 %

3.2 The Results of the Classification Based on Feature Subsets

The XGBoost is used for binary classification of the anomalous traffic into normal and then anomalous traffic. The feature set used in each instance and the performance attained by the XGBoost model are shown in Table 6, Figure 6 and Figure 7. The results showed that the highest accuracy of 99.85 % occurred when 60 or more features are used. At the same time, the model training time also is at its highest value of 53 seconds when the 60 or more features are used.

Table 6 The performance of the XGBoost classifier for the six different feature subsets used.

No. of Features	Accuracy (%) for XGBoost	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)	Training Time (seconds) for XGBoost	ROC AUC
10	0.9971	1	1	1	1.99	0.0001
20	0.998	1	1	1	2.76	0.0001
30	0.9982	1	1	1	3.47	0.0001
40	0.9983	1	1	1	4.34	0.0001
50	0.9984	1	1	1	5.37	0.0001
60	0.9985	1	1	1	6.53	0

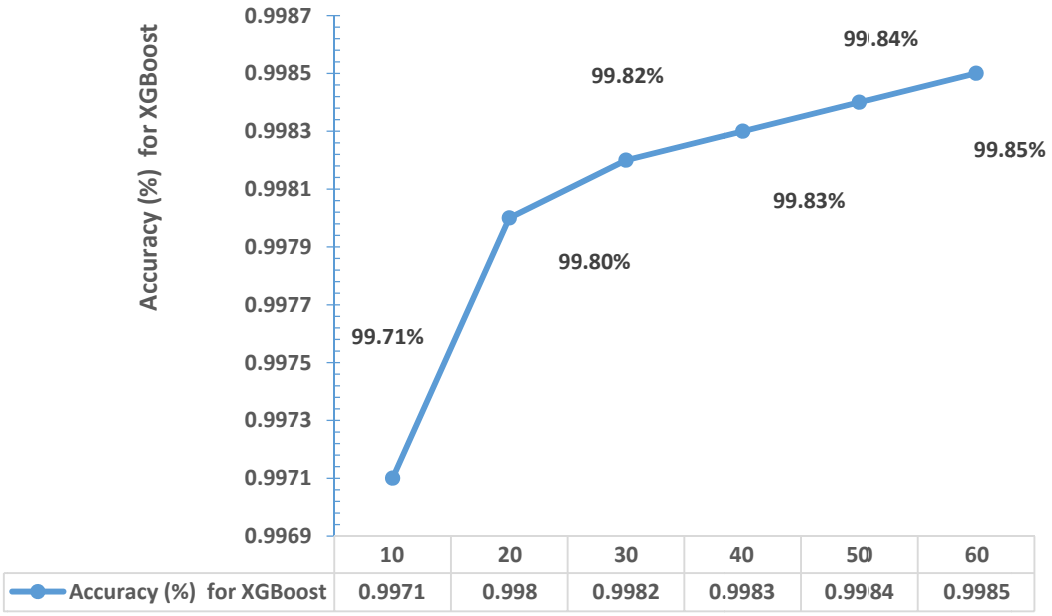

Figure 6 The accuracy for XGBoost for the multiclass classification using selected features

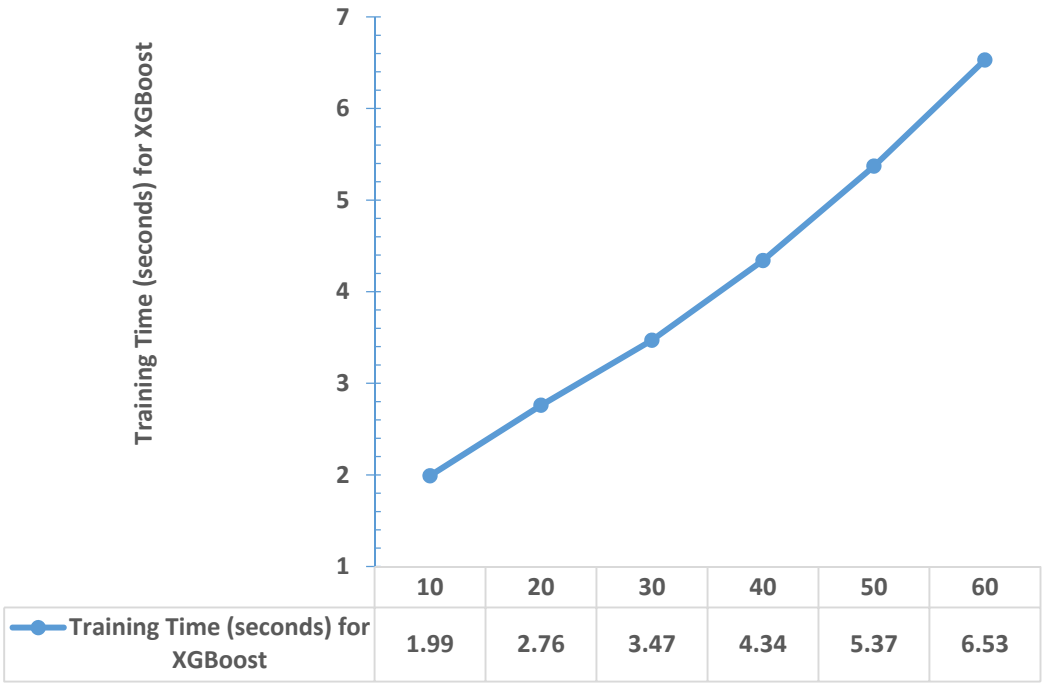

Figure 7 The training time for XGBoost for the multiclass classification using selected features

In order to determine the appropriate number of feature to be recommended for the XGBoost model, the variation or sensitivity of the accuracy and training time with number of features are plotted in Figure 8 and Figure 9. The plot in Figure 8 shows that after about 22 feature, the rate of increase in accuracy for every new feature added drops and keeps dropping till the all the features are added. On the other hand, the plot in Figure 9 shows that after about 22 feature, the rate of increase in accuracy for every new feature added decreased but it keeps increasing till the all the features are added. In essence, the first 20 to 22 features can be used without losing much accuracy but saving the execution time.

At exactly 20 features, the model accuracy is 99.8 % with training time of 2.76 seconds. On the other hand, at the 60^{th} feature, the model accuracy is 99.88 % with training time of 6.53 seconds. This shows that system resources can be preserved by adopting 20 feature subset without losing any significant value in the accuracy of the model.

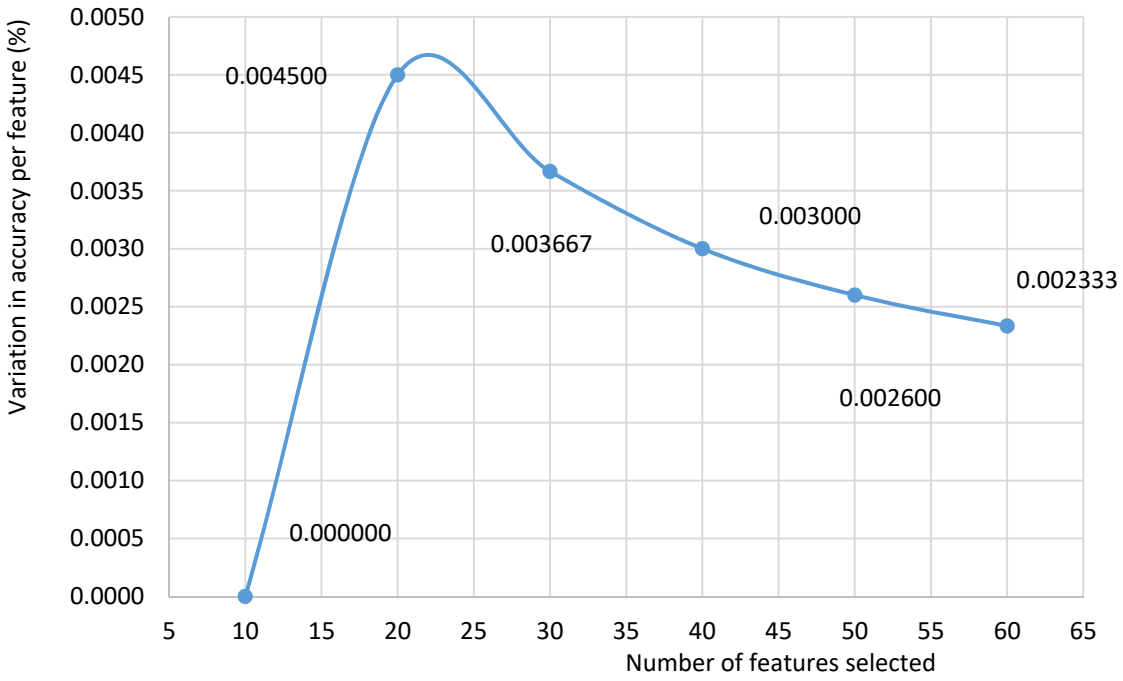

Figure 8 Variation in model training per feature selected for the XGBoost model

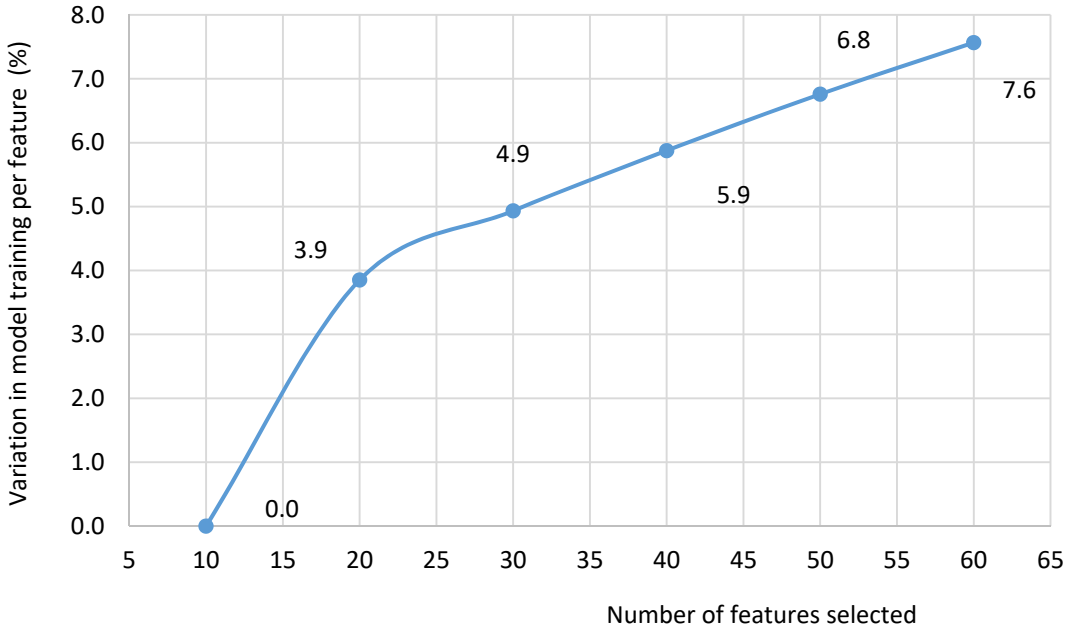

Figure 9 Variation in model training time per feature selected for the XGBoost model

4. Conclusion

The XGBboost model is presented for binary classification using various subsets of the features of an Industrial Internet of things Intrusion Detection (IIoTID) dataset. The study focus is on evaluation of the effect of feature subset on the performance of the XGBoost classifier. Specifically, the accuracy and training time of the model are used to evaluate the model and to identify the appropriate threshold number of feature to be recommended for the given dataset training and validation. The results show that the accuracy of the XGboost model continue to increase as the number of features increases. However, the rate of increase in the accuracy and the training time showed that the accuracy sensitivity drops after 22 features whereas, the sensitivity of the training time continues to increase with addition of new features. In all, the feature subset consisting of the top 20 to 22 features is recommended as the most appropriate for the case study XGBoost model and IIoTID dataset.

References

1. Sarker, Iqbal H. "AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems." *SN computer science* 3.2 (2022): 158.
2. Rayhan, Abu. "Artificial intelligence in robotics: From automation to autonomous systems." *IEEE Transactions on Robotics* 39.7 (2023): 2241-2253.
3. Bathla, Gourav, et al. "Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities." *Mobile Information Systems* 2022.1 (2022): 7632892.
4. Such applications are seen in autonomous vehicles, smart grids, smart agriculture and smart manufacturing systems
5. Sahoo, Snehasis, and Cheng-Yao Lo. "Smart manufacturing powered by recent technological advancements: A review." *Journal of Manufacturing Systems* 64 (2022): 236-250.

6. Khalifeh, Ala, et al. "Microcontroller unit-based wireless sensor network nodes: A review." *Sensors* 22.22 (2022): 8937.
7. Nourildean, Shayma Wail, Mustafa Dhia Hassib, and Y. A. Mohammed. "Internet of things based wireless sensor network: a review." *Indones. J. Electr. Eng. Comput. Sci* 27.1 (2022): 246-261.
8. Ahmed, Shams Forruque, et al. "Industrial Internet of Things enabled technologies, challenges, and future directions." *Computers and Electrical Engineering* 110 (2023): 108847.
9. Peter, Onu, Anup Pradhan, and Charles Mbohwa. "Industrial internet of things (IIoT): opportunities, challenges, and requirements in manufacturing businesses in emerging economies." *Procedia Computer Science* 217 (2023): 856-865.
10. Hu, Yujiao, et al. "Industrial internet of things intelligence empowering smart manufacturing: A literature review." *IEEE Internet of Things Journal* 11.11 (2024): 19143-19167.
11. Arents, Janis, and Modris Greitans. "Smart industrial robot control trends, challenges and opportunities within manufacturing." *Applied Sciences* 12.2 (2022): 937.
12. Chi, Hao Ran, et al. "A survey of network automation for industrial internet-of-things toward industry 5.0." *IEEE Transactions on Industrial Informatics* 19.2 (2022): 2065-2077.
13. Khan, Izhar Ahmed, et al. "Enhancing IIoT networks protection: A robust security model for attack detection in Internet Industrial Control Systems." *Ad Hoc Networks* 134 (2022): 102930.
14. Alabadi, Montdher, Adib Habbal, and Xian Wei. "Industrial internet of things: Requirements, architecture, challenges, and future research directions." *IEEE Access* 10 (2022): 66374-66400.

15. Rozlomii, Inna, Andrii Yarmilko, and Serhii Naumenko. "Data security of IoT devices with limited resources: challenges and potential solutions." *doors* 3666 (2024): 85-96.
16. Alabadi, Montdher, Adib Habbal, and Xian Wei. "Industrial internet of things: Requirements, architecture, challenges, and future research directions." *IEEE Access* 10 (2022): 66374-66400.
17. Wang, Maoli, et al. "Security issues on industrial internet of things: Overview and challenges." *Computers* 12.12 (2023): 256.
18. Awotunde, Joseph Bamidele, Chinmay Chakraborty, and Abidemi Emmanuel Adeniyi. "Intrusion detection in industrial internet of things network-based on deep learning model with rule-based feature selection." *Wireless communications and mobile computing* 2021.1 (2021): 7154587.
19. Alkahtani, Hasan, and Theyazn HH Aldhyani. "Intrusion Detection System to Advance Internet of Things Infrastructure-Based Deep Learning Algorithms." *Complexity* 2021.1 (2021): 5579851.
20. Saheed, Yakub Kayode, et al. "A novel hybrid autoencoder and modified particle swarm optimization feature selection for intrusion detection in the internet of things network." *Frontiers in Computer Science* 5 (2023): 997159.