IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 January 2023, accepted 18 February 2023, date of publication 27 February 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3249663

RESEARCH ARTICLE

Classification and Prediction of Significant Cyber Incidents (SCI) Using Data Mining and Machine Learning (DM-ML)

GOHAR MUMTAZ[1,2], SHEERAZ AKRAM[1,2,3], MUHAMMAD WASEEM IQBAL[4], M. USMAN ASHRAF[5], KHALID ALI ALMARHABI[6], AHMED MOHAMMED ALGHAMDI[7], AND ADEL A. BAHADDAD[8]

[1]Faculty of Computer Science and Information Technology, Superior University, Lahore 54000, Pakistan
[2]Intelligent Data Visual Computing Research (IDVCR), Lahore 73861, Pakistan
[3]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11564, Saudi Arabia
[4]Department of Software Engineering, Faculty of Computer Science and Information Technology, Superior University, Lahore 54000, Pakistan
[5]Department of Computer Science, GC Women University, Sialkot, Sialkot 51310, Pakistan
[6]Department of Computer Science, College of Computing in Al-Qunfudah, Umm Al-Qura University, Mecca 21421, Saudi Arabia
[7]Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia
[8]Department of Information System, King Abdul Aziz University, Jeddah 21589, Saudi Arabia

Corresponding author: M. Usman Ashraf (usman.ashraf@gcwus.edu.pk)

This work was supported by the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, under Project IFP22UQU4310108DSR188.

ABSTRACT The rapid growth in technology and several IoT devices make cyberspace unsecure and eventually lead to Significant Cyber Incidents (SCI). Cyber Security is a technique that protects systems over the internet from SCI. Data Mining and Machine Learning (DM-ML) play an important role in Cyber Security in the prediction, prevention, and detection of SCI. This study sheds light on the importance of Cyber Security as well as the impact of COVID-19 on cyber security. The dataset (SCI as per the report of the Center for Strategic and International Studies (CSIS)) is divided into two subsets (pre-pandemic SCI and post-pandemic SCI). Data Mining (DM) techniques are used for feature extraction and well know ML classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) for classification. A centralized classifier approach is used to maintain a single centralized dataset by taking inputs from six continents of the world. The results of the pre-pandemic and post-pandemic datasets are compared and finally conclude this paper with better accuracy and the prediction of which type of SCI can occur in which part of the world. It is concluded that SVM and RF are much better classifiers than others and Asia is predicted to be the most affected continent by SCI.

INDEX TERMS Significant cyber incidents, cyber security, data mining, machine learning.

I. INTRODUCTION

The speedy advancement in technology and the boom in the IoT industry increase the possibility of cyber incidents. Especially, after the pandemic COVID-19, this ratio is in-creased [1], [2]. It is expected that the number of IoT devices count will reach around 75 billion by 2025 [3]. As per the handbook 'Cybersecurity Almanac' released by 'Cybersecurity Ventures, the global cybercrime cost is expected to reach USD 10.5 trillion in 2025, from USD 6 trillion in 2021. In 2021, an organization suffered from a ransomware attack after every 11 seconds, and it is expected to suffer after every 2 seconds in 2031 [4]. Table 1 depicts up-to-date statistics about the internet and social media users from January 2020 to October 2022. There is an alarming increase in the percentage of 24.5 active social media users [5].

Cyber security is a technique to protect systems over the internet from cyber incidents. A cyber incident means an activity or event which occurred through the internet

The associate editor coordinating the review of this manuscript and approving it for publication was Chien-Ming Chen.

TABLE 1. Internet and social media users' statistics.

Stats	*January 2020*	*October 2022*	*Increase %*
Total Population	7.75 billion	7.99 billion	3.1
Unique Mobile Users	5.19 billion	5.48 billion	5.6
Internet Users	4.54 billion	5.07 billion	11.7
Active Social-Media Users	3.80 billion	4.74 billion	24.5
Average Internet user time each day	6 Hours 43 Minutes	7 Hours	5

and jeopardizes the Confidentiality, Integrity, and Availability (CIA Triad) of the communication system through any means [6]. The term Significant Cyber Incident (SCI) means a cyber incident that results in manifest damage to the national security and economy [7]. Cyber security is used by individuals as well as organizations to protect their information and systems over the internet from unauthorized access.

With the increase in SCI, cyber security measures also improved to tackle these incidents. Data Mining and Machine Learning (DM-ML) play an important role in cyber incidents prediction, prevention, and Detection by using different approaches [8], [9], [10].

In this paper, the outfall of SCI has been predicted based on the datasets, collected from the report of the Center for Strategic and International Studies (CSIS) [11]. The datasets consist of textual data comprising of SCI type and the continent where it occurred. First, it is divided into two parts (pre-pandemic SCI and post-pandemic SCI) and then analyzed ten types of SCI that occurred in six continents of the world. Pre-pandemic (before COVID-19) dataset includes those SCI which happened during the period from 2003 to December 2019. Similarly, the post-pandemic (after COVID-19) dataset includes those SCI which happened during the period from January 2020 to till date. As there are no countries in the seventh continent 'Antarctica', so the only six continents in our study are considered. Further, it is also investigated how the data can be used for classification accuracy and eventually the better classifier for distinguishing different SCI. The results achieved by focusing on which type of SCI occurred at which continent of the world. The main objective of this study is to explore the benefits of centralized classifier for treating future SCI.

Data Mining features like n-grams and Bag of Words (BoW) are more useful now for the feature extraction from the collected data [12], [13], [14]. ML algorithms like Naïve Bayes (NB) [15], [16], Support Vector Machine (SVM) [17], [18], Logistic Regression (LR) [19], [20] and Random Forest (RF) [21], [22] are used for data classification [23]. Finally, the results of pre- and post-pandemic datasets are compared which concludes with the best results of SVM, and RF classifiers and Asia (the most affected continent by SCI) is predicted.

In Section II, a detailed review of DM-ML approaches to combat these SCI is discussed. In Section III, a methodology model used to get optimized results is proposed. In Section IV, the output of the different classifiers is compared and finally in Section V, it is concluded that SVM and RF are both better than others.

II. LITERATURE REVIEW

Cyber security is an emerging and enormous challenge in the world and concerns dealing with different cyber incidents [1]. The important part is to identify the existing incidents using different DM-ML algorithms. DM helps to identify similarities in the cyber incidents' patterns while ML trains models and predicts cyber incidents. This re-view section presents existing DM-ML approaches to prevent and predict cyber incidents [24]. DM is concerned with hidden knowledge or pattern of the dataset while ML is concerned with training the system based on hidden knowledge. DM approaches like Association, Classification and Clustering help in identifying and recognizing the behaviors of cyber incidents. Association-based algorithms provide real-time prediction and prevention of cyber incidents because of the strong incident and vulnerability linkage [25]. The classification method classifies the existing dataset and is eventually very helpful in the prediction of Denial-of-Service (DoS) and Distributed Denial-of-Service (DDoS) cyber incidents [26]. In the DM Clustering technique, similar types of incidents are grouped, and this strategy helps in rectifying phishing cyber incidents [27].

The authors in [28] used a text mining approach to detect cyber incidents in digital healthcare. The authors used Natural Language Processing (NLP) to mine news data and get insight. Song and Suh [29], proposed a novel framework using text mining for the assessment and detection of cyber risk. In [30], The authors proposed an anomaly detector using accident reports. They worked on textual data and used the Local Outlier Factor (LoF) for anomalous condition detection. The authors surveyed different DM-ML approaches for malware detection [31], [32]. In another paper, the authors used a Deep learning methodology is used for forecasting cyber-attacks based on the captured data from network traffic [33], [34], [35]. In another paper [36], cyber-attack methods and committers have been predicted using Support Vector Machine (SVM), an ML algorithm. In [37], the authors concluded different DM-ML approaches like Bayesian network, Decision Tree, Clustering, and Artificial Neural Networks (ANN) in cyber security to detect cyber incidents.

The main motivation of this research study is a centralized classifier that collects data from the six continents of the world. This approach maintains a centralized dataset, comprising efforts from six continents. Further, DM-ML-based approaches are very well-known techniques being used to detect vulnerability in cyber security and that is the reason n-gram and BoW are used for feature extraction in the model. For the classifier, NB, SVM, LR and RF are used.

TABLE 2. Distribution of the pre-pandemic (before COVID-19) dataset.

SCI Type	Africa	Asia	Europe	North America	Oceania	South America	*SCI Type Count*
APT	1	21	12	19	2	1	56
Brute Force	0	0	1	1	0	0	2
DDoS	1	11	7	11	1	0	31
DoS	0	18	11	15	2	0	46
Espionage	0	16	24	27	2	0	69
Malware	0	44	21	48	1	1	115
Man-in-Middle	0	7	7	8	1	0	23
Phishing	3	35	21	43	1	4	107
SQL Injection	1	5	3	16	1	0	26
Zero-Day Exploit	0	10	9	9	1	0	29
Total	6	167	116	197	12	6	504

TABLE 3. Distribution of the post-pandemic (after COVID-19) dataset.

SCI Type	Africa	Asia	Europe	North America	Oceania	South America	*SCI Type Count*
APT	4	62	25	20	5	0	116
DDoS	0	15	13	6	3	0	37
DoS	0	1	0	0	0	0	1
Espionage	1	8	7	7	1	0	24
Malware	4	46	53	19	3	0	125
Man-in-Middle	1	1	1	1	1	0	5
Phishing	3	55	86	47	6	8	205
SQL Injection	2	3	6	2	0	0	13
Zero-Day Exploit	0	12	3	1	1	0	17
Total	15	203	194	103	20	8	543

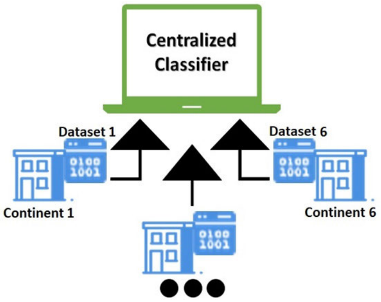


FIGURE 1. Centralized classifier for data collection from six continents.

III. METHODOLOGY

This study is proposed to classify SCI using DM-ML approaches. The main objective in this study is the centralized classifier, as shown in Fig. 1. The dataset is collected from the six continents of the world.

The dataset is used to train the centralized classifier and eventually a better performance rather than using a separate classifier for each continent.

A. DATASET INTERPRETATION

The dataset is the type of SCI, that occurred in 6 continents of the world (from September 2003 to October 2022), as per the report of the Center for Strategic and International Studies (CSIS) [11]. The dataset is classified as continent wise and there is a total of 1047 SCI. Further, it is divided into two parts (pre-pandemic and post-pandemic) for comparative analysis. The distributions of the datasets are shown in Table 2 and Table 3. There are 504 SCI in the pre-pandemic dataset and 543 SCI in the post-pandemic dataset. The SCI count for Asia is higher because it is the largest continent in the world (See Fig. 2.). The impact of COVID-19 on SCI (for the year 2020 and year 2021) can be seen in Fig. 3. Each SCI entry (in a single row) has the following details. From the dataset, SCI type is chosen as labeled data.

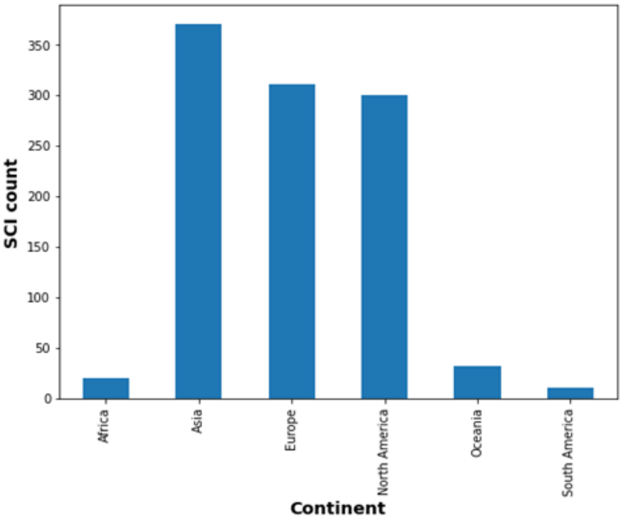


FIGURE 2. Continent wise SCI count.

- Name of the Continent (e.g., Asia)
- Name of the Country (e.g., Pakistan)

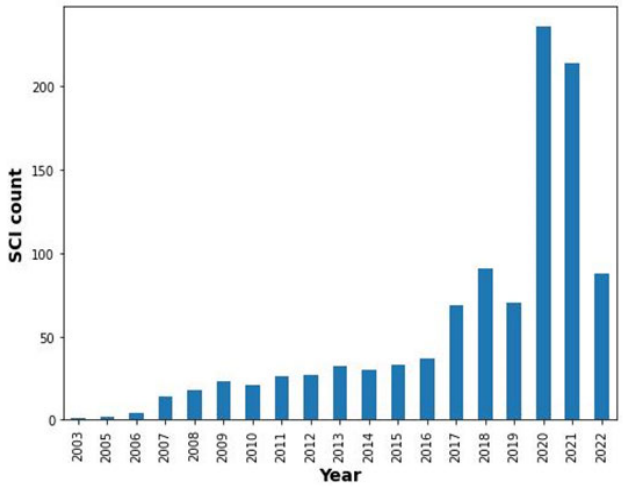


FIGURE 3. Year wise SCI count.

- Year (e.g., 2021)
- Month (e.g., January)
- SCI Type (e.g., DDoS)

B. RESEARH PROBLEM

This study identified, investigated, and solved how to forecast the name of the continent based on the type of SCI. For this, four types of classifiers are used to check the efficiency of the classifiers for the collected datasets.

C. METHOD OF ANALYSIS AND CLASSIFICATION

This section is further divided into three sub-sections. The whole process is also shown in Algorithm 1.

1) DATA COLLECTION AND PRE-PROCESSING

The data is collected from the report of the Center for Strategic and International Studies (CSIS) and this report is updated monthly. The aim of the pre-processing data is to remove noise. The results are improved by reducing errors in the dataset. For this, extra words are removed and sorted out ambiguities due to null values in the dataset as well. Then a single column (Continent) is chosen with labeled data (SCI type) to solve our research problem. At this stage, our dataset is ready for feature extraction.

2) FEATURE EXTRACTION

For feature extraction, the most common features of data mining are n-gram and BoW. We used the concept of uni-gram and bi-gram models. For uni-gram, n=1, for example, "phishing" is a word, and all the SCI are extracted containing this word from the dataset. And for bi-gram, n=2, for example, "SQL Injection" are two words, and the SCI are extracted based on these two adjacent words. At this stage, BoW is used to carry all uni-gram and bi-gram words. This BoW is filtered to filter out words with minimum frequency of their occurrence in the dataset. These words are not further used for features using Term Frequency – Inverse Document Frequency (TF-IDF), a technique based on the BoW model is used for text vectorization (in our case feature vector (504, 6) for pre-pandemic dataset and (543,6) for post-pandemic dataset).

Algorithm 1 Data Analysis and Classification

```
Variables:   P, X, n, Z, S, N, L, R
             P: data after pre-processing
             X: data after feature extraction
             n: number of words
             Z: feature Vector
             S: SVM Classifier
             N: Naïve Bayes Classifier
             L: Logistic Regression Classifier
             R: Random Forest Classifier
Input (I):   [I: Unclassified instances]
Output (Ω):  [Ω: Classified instances]
Procedure:
             pre-processing [removed null values]
             feature extraction for words (n)
             If n(1)        uni-gram
             If n(2)        bi-gram
Classification:
               Classifiers: S, N, L, R
               For Classifiers: Input (Z)
               Class APT to Zero-Day Exploit
                   Case 1:            Asia
                   Case 2:            Africa
                   Case 3:            Europe
                   Case 4:     North America
                   Case 5:     South America
                   Case 6:           Oceania
             Return (Ω) Classified Instances.
```

3) CLASSIFICATION

Four different classifiers of Machine learning are used one by one.

a: NAÏVE BAYES (NB)

It is based on Bayes' Theorem, derived from conditional probability. It is commonly used in supervised learning for text data classification. NB is efficient for nonlinear problems because of non-biasedness by outliers and is not efficient if the assumptions are based on statistically relevant features. When it comes to classification, NB is a form of probabilistic learning, so it is used for categorizing texts. One of the most well-known algorithms, it is used to categorize documents into one or more groups [38], [39].

b: SUPPORT VECTOR MACHINE (SVM)

It is a soft margin classifier, and it is commonly used to detect outliers. It is also a supervised learning classifier. SVM is a vector-on-point approach and is very efficient if the problem is linear and the dataset is limited. It is also

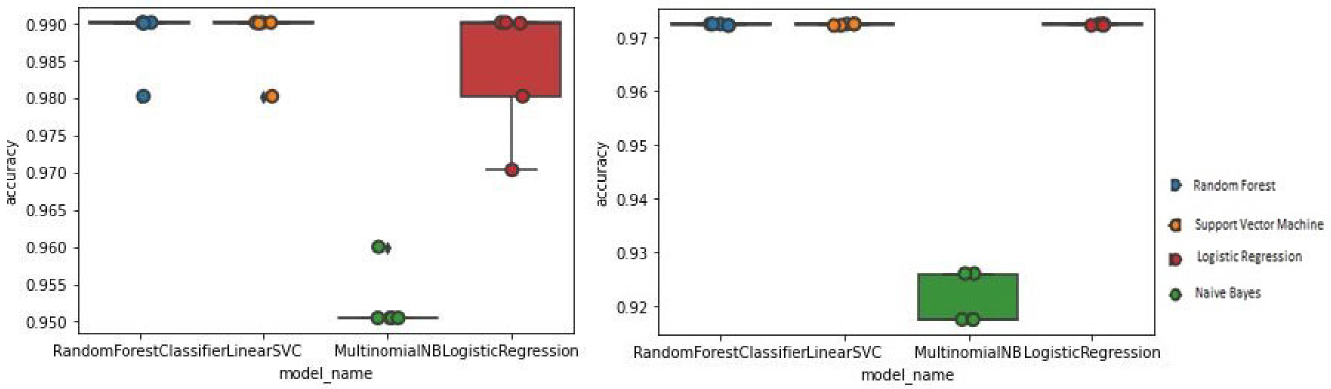


FIGURE 4. Accuracy measure for different classifiers (Left: pre-pandemic dataset. Right: post-pandemic dataset).

good for nonlinear problems and datasets with many features. It has become the standard for cutting-edge machine learning applications. In machine learning, support-vector machines are supervised learning models that analyze data for classification and regression using corresponding learning methods.

SVMs may easily do a non-linear classification in addition to their typical linear classification by utilizing the kernel technique to implicitly transform their inputs into high-dimensional feature spaces [40]. Applications where SVMs are employed include web page classification [41], email classification [42], intrusion detection [43], face identification, and handwriting recognition [44].

c: LOGISTIC REGRESSION (LR)

Logistic Regression predicts the binary problem and its outcome efficiently. It gives information about the statistical signification of features and uses a Probabilistic approach. Its efficiency can be increased by normalizing the data. Logistic Regression is a Straight Line, Logarithmic Line approach. The logistic sigmoid function is used to provide a probability value as a transform in logistic regression. In classification issues like determining if an email is spam or not [45], or whether a tumor is malignant or benign [46], logistic regression is used as a classification procedure to assign data to a discrete set of classes.

d: RANDOM FOREST (RF)

Random Forest uses an ensemble learning technique. It consists of many decision trees and by increasing the number of trees, the efficiency of the model also increases. It also works on nonlinear problems. In technical terms, it is an ensemble method (using a divide-and-conquer strategy) for generating decision trees from a subset of a dataset.

The collection of decision trees used as classifiers is also known as a random forest [47]. In a classification problem, each tree acts as a vote, and the winning class is determined by the total number of votes it receives. RF excels at classification and regression problems where many entries and features (often with missing values) are present, helping to produce a highly accurate result while avoiding overfitting. Additionally, RF helps in revealing the relative feature importance, letting to select the most important features.

In this study, these four classifiers are used as our dataset is non-linear Because of the independence of its features and we want to explore the ability of the classifier.

IV. EXPERIMENTS AND RESULTS

This research focuses on four different types of classifiers to perform an experimental study to explore the ability of the classifier. The output of the classifier is to predict the name of the continent based on the type of SCI. To evaluate the performance of the classifiers, we used Accuracy, Recall, Precision and F1-measure as performance indicators.

To evaluate the performance of the classifiers, the model is trained through 1047 SCI. The Accuracy measure after training all the four classifiers is shown in Fig. 4. The results clearly show that the accuracy for SVM classifier after the training (0.988099 for pre-pandemic dataset and 0.972375 for post-pandemic dataset) is higher than the others.

The accuracy measures (pre-pandemic dataset, post-pandemic dataset) for NB, LR, and RF are (0.952396, 0.920829), (0.984139, 0.962375), (0.978099, 0.962375) respectively. Further, the model is tested and predicted the output based on the type of SCI. Each classifier is tested one by one and compared with the previous one. In the end, the efficient classifier is concluded, particularly for the case under consideration.

Firstly, SVM classifier is used to predict the output. The evaluated results are shown in Table 4, while the confusion matrix for the SVM model is shown in Fig. 5. The Accuracy of this model is 99% and 96% for pre-pandemic and post-pandemic cases respectively. The values of Precision, Recall and F1 measure against Africa are all approximately 0.02 for

TABLE 4. Models evaluated using SVM.

	Pre-Pandemic				*Post-Pandemic*			
Average Accuracy	0.99				0.96			
Class	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**Accuracy**	**Precision**	**Recall**	**F1-measure**
Asia	0.98	0.98	0.99	0.99	0.97	0.98	0.99	0.97
North America	0.99	0.98	0.96	0.96	0.96	0.97	0.96	0.99
Europe	0.98	0.97	0.98	0.97	0.97	0.99	0.98	0.98
Oceania	0.99	0.67	0.98	0.80	0.96	0.42	0.97	0.59
Africa	0.98	0.05	0.08	0.04	0.95	0.02	0.05	0.07
South America	0.99	0.98	0.99	0.96	0.96	0.99	0.96	0.98

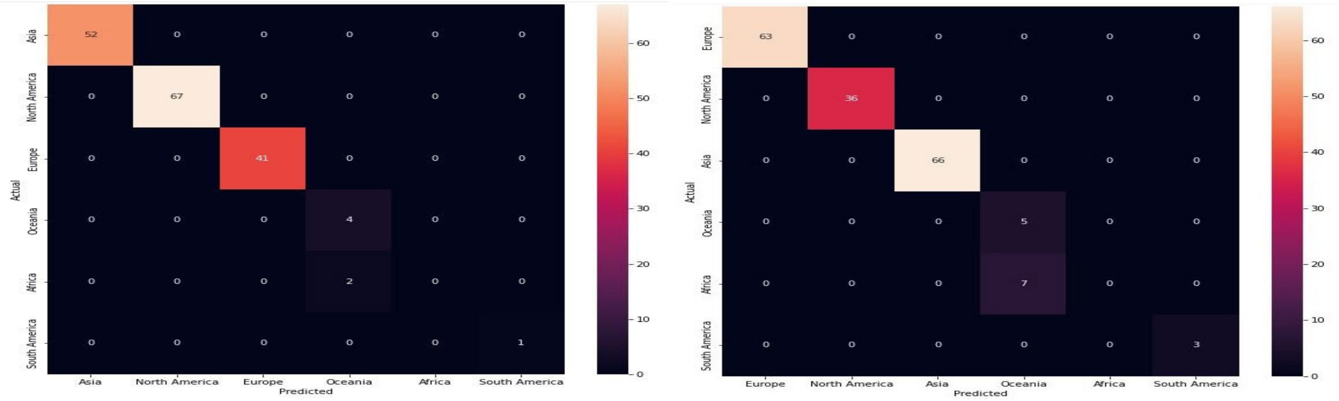


FIGURE 5. SVM confusion matrix (Left: pre-pandemic dataset. Right: post-pandemic dataset).

both cases because of a very limited number of SCI in this continent.

The precision, Recall and F1 measures against Asia, Europe, North America, and South America, are all approximately 0.99 which is quite good and shows the goodness of the classifier. Different evaluation indicators are used which predict the accuracy of the study. The same results are observed in the Confusion matrix which identifies actual and predicted results, as shown in Fig. 5. The SVM classifier confusion matrix clearly shows the actual values and predicted values against each continent. For instance, the values of the Asia continent showed 52 for a pre-pandemic case and 66 for a post-pandemic case for the actual and the same values for the predicted, which shows the precision values in Table 3. All the other continent values can also be compared and checked in the evaluations measure. The heatmap in the study shows the accuracy level against actual and predicted values which is shown in Fig. 5.

Secondly, NB Classifier is used to check its efficiency against other models. The evaluated result against this classifier is shown in Table 5 and the confusion matrix is shown in Fig. 6. It is clear from the results that this classifier is not better than SVM. Its accuracy is 96% for the pre-pandemic cases and 92% for the post-pandemic cases, which is too less than SVM and the values of precision, recall and f1-measure are approximately 0.02 against three continents.

From the confusion matrix, we see Asia-Asia cross values for both pre-pandemic as well as the post-pandemic cases, which are quite good in terms of actual vs predicted. This shows the efficiency of the model only for the Asia continent. The Oceania-North America cross value is 4 in pre-pandemic case, and Africa-Asia cross value is 7 due to which its accuracy level goes down. Here again, the focus is on the Asia continent due to the large sample size in the dataset. It is concluded here that NB is not better than SVM.

Thirdly, LR classifier is used, and the values for Precision, Recall and F1-measure against Africa are approximately 0.01. As compared with SVM, the accuracy of LR is less than SVM and This is the main reason SVM is better than LR. The detailed evaluated result is depicted in Table 6 and the confusion matrix in Fig. 7.

The actual versus predicted results are quite good against Asia, North America, and South America as shown by

TABLE 5. Models evaluated using NB.

	Pre-Pandemic				*Post-Pandemic*			
Average Accuracy	0.96				0.92			
Class	Accuracy	Precision	Recall	F1-measure	Accuracy	Precision	Recall	F1-measure
Asia	0.98	0.99	0.99	0.98	0.93	0.97	0.96	0.98
North America	0.96	0.91	0.99	0.95	0.92	0.92	0.98	0.96
Europe	0.97	0.98	0.97	0.99	0.91	0.85	0.99	0.92
Oceania	0.96	0.02	0.01	0.04	0.94	0.02	0.03	0.04
Africa	0.95	0.01	0.02	0.03	0.92	0.01	0.04	0.03
South America	0.97	0.04	0.03	0.01	0.91	0.01	0.03	0.04

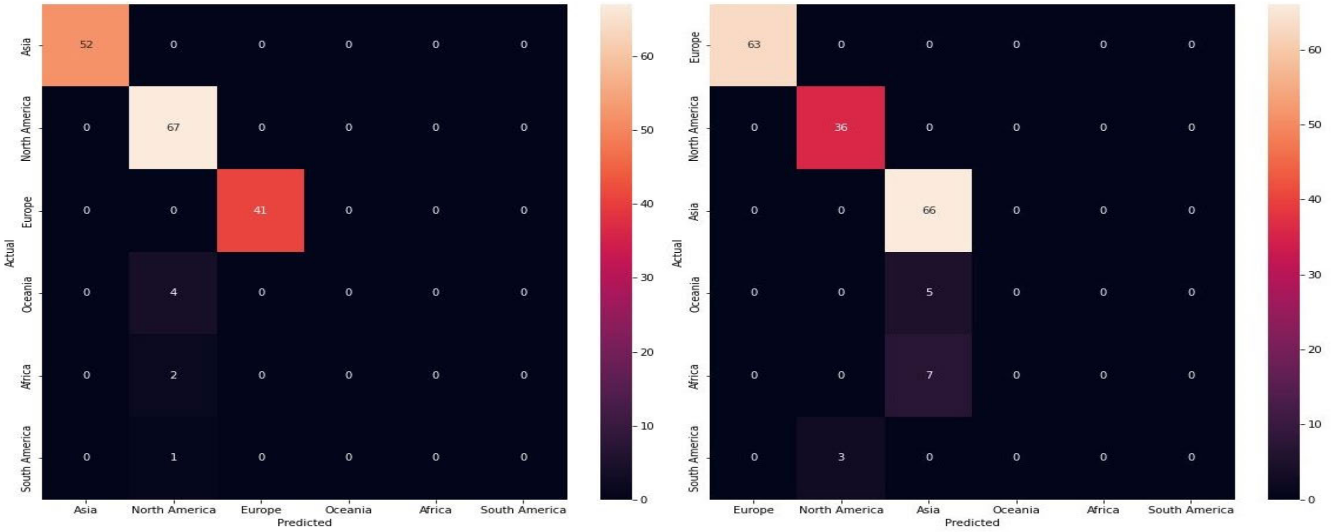


FIGURE 6. NB confusion matrix (Left: pre-pandemic dataset. Right: post-pandemic dataset).

TABLE 6. Models evaluated using LR.

	Pre-Pandemic				*Post-Pandemic*			
Average Accuracy	0.98				0.95			
Class	Accuracy	Precision	Recall	F1-measure	Accuracy	Precision	Recall	F1-measure
Asia	0.99	0.99	0.98	0.98	0.97	0.99	0.98	0.97
North America	0.97	0.98	0.97	0.99	0.95	0.96	0.98	0.99
Europe	0.98	0.96	0.99	0.98	0.95	0.99	0.97	0.99
Oceania	0.97	0.57	0.99	0.70	0.95	0.32	0.98	0.49
Africa	0.96	0.01	0.03	0.02	0.94	0.01	0.04	0.03
South America	0.98	0.98	0.99	0.98	0.96	0.97	0.98	0.99

confusion matrix. This concludes LR classifier is a good classifier but when compared to SVM, The SVM is better than LR. In the confusion matrix (post-pandemic case), the Africa-Oceania cross value is 7 which means there were 7 such SCI (occurred in Africa) and the classifier predicted Oceania which is the wrong prediction. This is

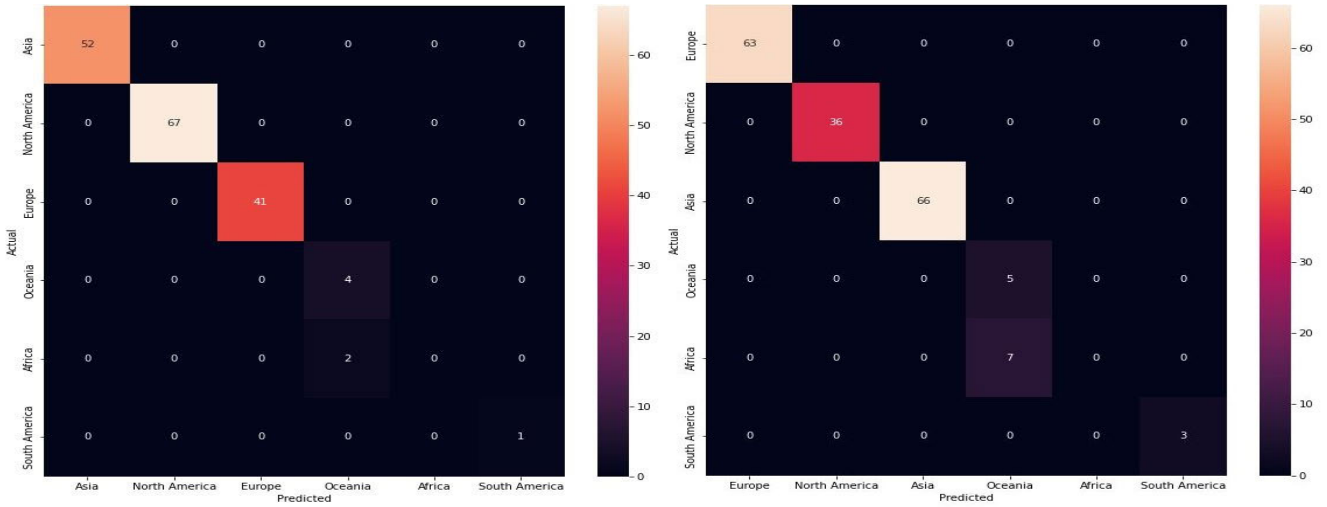


FIGURE 7. LR confusion matrix (Left: pre-pandemic dataset. Right: post-pandemic dataset).

TABLE 7. Models evaluated using RF.

	Pre-Pandemic				*Post-Pandemic*			
Average Accuracy	**0.99**				**0.96**			
Class	**Accuracy**	**Precision**	**Recall**	**F1-measure**	**Accuracy**	**Precision**	**Recall**	**F1-measure**
Asia	0.99	0.99	0.98	0.97	0.98	0.98	0.98	0.97
North America	0.98	0.98	0.99	0.96	0.96	0.97	0.98	0.99
Europe	0.98	0.99	0.98	0.99	0.95	0.97	0.99	0.98
Oceania	0.99	0.67	0.99	0.80	0.96	0.42	0.99	0.59
Africa	0.99	0.01	0.03	0.04	0.95	0.02	0.03	0.04
South America	0.98	0.99	0.98	0.97	0.97	0.98	0.99	0.97

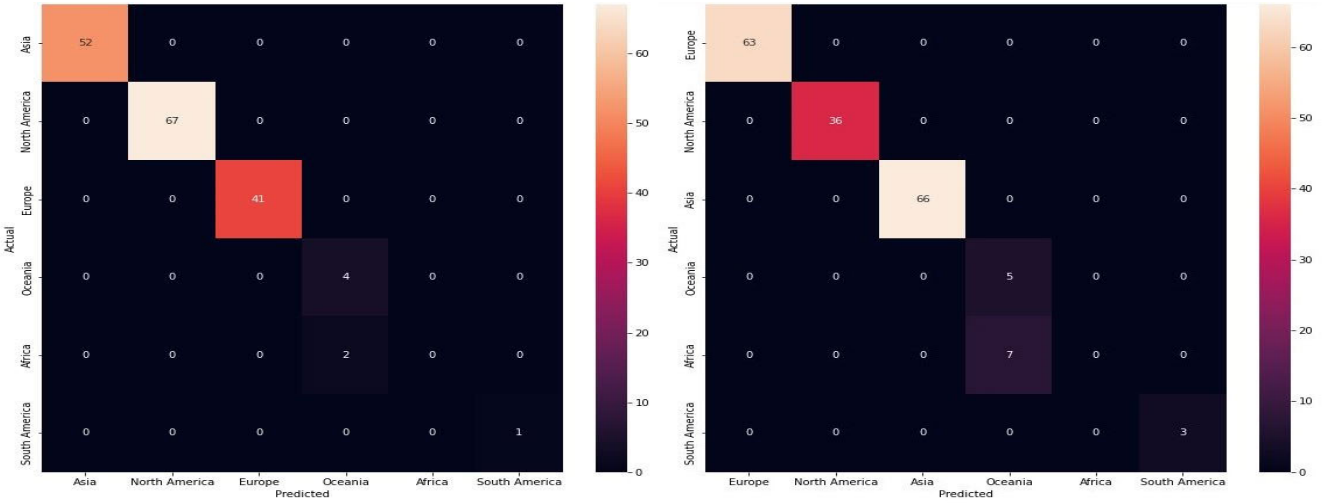


FIGURE 8. RF confusion matrix (Left: pre-pandemic dataset. Right: post-pandemic dataset).

one of the reasons the accuracy level goes down when LR is used.

Finally, and fourthly, the RF classifier is used and evaluated results with the accuracy of 99% (pre-pandemic) and 96%

(post-pandemic) are shown in Table 7. The confusion matrix is shown in Fig. 8. The results are exactly like SVM. One reason for the similar result is the non-linearity of the data and limited dataset.

SVM performs very well for the limited dataset. RF results efficiently for imbalanced classes. SVM, NB, LR and RF classifiers are evaluated one by one. It is concluded that SVM and RF are both good classifiers for the case under consideration.

V. CONCLUSION AND FUTURE WORK

This paper focuses on the research based on Significant cyber incidents (SCI) from September 2003 to October 2022 as per the report of the Center for strategic and international studies (CSIS). The datasets are analyzed and classified using data mining and machine learning algorithms. Four different classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) are used and predicted the output (name of the continent based on the type of SCI). It is also predicted which continent is more affected by SCI during the period. Finally, it is concluded that SVM and RF are both better than other classifiers against our models, in both cases (pre-pandemic and post-pandemic) and Asia is the most affected continent by SCI.

In the future, different datasets can be considered for SCI and can apply different classifiers with advanced machine learning techniques like Federated Machine Learning (FML) to check the efficiency of the classifiers. Further, Blockchain can also be implemented in our model to enhance security.

REFERENCES

[1] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security; emerging trends and recent developments," *Energy Rep.*, vol. 7, pp. 8176–8186, Nov. 2021, doi: 10.1016/j.egyr.2021.08.126.

[2] J. Kaur and K. R. Ramkumar, "The recent trends in cyber security: A review," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5766–5781, Sep. 2022, doi: 10.1016/j.jksuci.2021.01.018.

[3] H. Hejase, H. Kazan, A. Hejase, and I. Moukadem, "Cyber security paper," *Comput. Inf. Sci.*, vol. 14, pp. 10–25, Mar. 2021, doi: 10.5539/cis.v14n2p10.

[4] *Cybersecurity Almanac by Cyber Security Ventures*. [Online]. Available: https://cybersecurityventures.com/cybersecurity-almanac-2022

[5] *Digital 2022 October Global Statshot, by Datareportal*. [Online]. Available: https://datareportal.com/

[6] P. S. Seemma, S. Nandhini, and M. Sowmiya, "Overview of cyber security," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 7, no. 11, pp. 125–128, Nov. 2018, doi: 10.17148/IJARCCE.2018.71127.

[7] Q. E. Hodgson, A. Clark-Ginsberg, Z. Haldeman, A. Lauland, and I. Mitch, *Managing Response to Significant Cyber Incidents: Comparing Event Life Cycles and Incident Response Across Cyber and Non-Cyber Events*. anta Monica, CA, USA: RAND Corp., 2022, doi: 10.7249/RRA1265-4.

[8] A. Handa, A. Sharma, and S. K. Shukla, "Machine learning in cybersecurity: A review," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 4, p. e1306, Jul. 2019, doi: 10.1002/widm.1306.

[9] A. E. Ibor, F. A. Oladeji, and O. B. Okunoye, "A survey of cyber security approaches for attack detection, prediction, and prevention," *Int. J. Secur. Appl.*, vol. 12, no. 4, pp. 15–28, Jul. 2018, doi: 10.14257/ijsia.2018.12.4.02.

[10] K. Shaukat Dar, S. Luo, S. Chen, and D. Liu, "Cyber threat detection using machine learning techniques: A performance evaluation perspective," in *Proc. Int. Conf. Cyber Warfare Secur. (ICCWS)*, Oct. 2020, pp. 1–6, doi: 10.1109/ICCWS48432.2020.9292388.

[11] *Significant Cyber Incidents (SCIs)*. [Online]. Available: https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents

[12] A. Pektaş, M. Eris, and T. Acarman, "Proposal of n-gram based algorithm for malware classification," in *Proc. 5th Int. Conf. Emerg. Secur. Inf., Syst. Technol.*, Jan. 2011, pp. 14–18.

[13] C. Wressnegger, G. Schwenk, D. Arp, and K. Rieck, "A close look on n-grams in intrusion detection: Anomaly detection vs. classification," in *Proc. ACM workshop Artif. Intell. Secur.*, Nov. 2013, pp. 14–18, doi: 10.1145/2517312.2517316.

[14] S. Soni and B. Bhushan, "Use of machine learning algorithms for designing efficient cyber security solutions," in *Proc. 2nd Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT)*, vol. 1, Jul. 2019, pp. 1496–1501, doi: 10.1109/ICICICT46008.2019.8993253.

[15] H. Alqahtani, I. H. Sarker, A. Kalim, S. M. M. Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques," in *Proc. Int. Conf. Comput. Sci., Commun. Secur.*, Singapore, 2020, pp. 121–131.

[16] A. Terai, S. Abe, S. Kojima, Y. Takano, and I. Koshijima, "Cyber-attack detection for industrial control system monitoring with support vector machine based on communication profile," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops*, Apr. 2017, pp. 132–138, doi: 10.1109/EuroSPW.2017.62.

[17] K. Ghanem, F. J. Aparicio-Navarro, K. G. Kyriakopoulos, S. Lambotharan, and J. A. Chambers, "Support vector machine for network intrusion and cyber-attack detection," in *Proc. Sensor Signal Process. Defense Conf. (SSPD)*, Dec. 2017, pp. 1–5, doi: 10.1109/SSPD.2017.8233268.

[18] N. Bhusal, M. Gautam, and M. Benidris, "Detection of cyber attacks on voltage regulation in distribution systems using machine learning," *IEEE Access*, vol. 9, pp. 40402–40416, 2021, doi: 10.1109/ACCESS.2021.3064689.

[19] R. Bapat, "Identifying malicious botnet traffic using logistic regression," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Apr. 2018, pp. 266–271, doi: 10.1109/SIEDS.2018.8374749.

[20] A. Kajal and G. Sardana, "Protection from cyber attacks using IDS security mechanism with random forest classifier: A review," *J. Crit. Rev.*, vol. 7, no. 19, p. 8516, 2020.

[21] S. Ustebay, Z. Turgut, and M. A. Aydin, "Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier," in *Proc. Congr. Big Data, Deep Learn. Fighting Cyber Terrorism*, Dec. 2018, pp. 71–76, doi: 10.1109/IBIGDELFT.2018.8625318.

[22] M. Malik, M. W. Iqbal, S. K. Shahzad, M. T. Mushtaq, M. R. Naqvi, M. Kamran, B. A. Khan, and M. Usman Tahir, "Determination of COVID-19 patients using machine learning algorithms," *Intell. Autom. Soft Comput.*, vol. 31, no. 1, pp. 207–222, 2022, doi: 10.32604/iasc.2022.018753.

[23] N. M. Chayal and N. P. Patel, "Review of machine learning and data mining methods to predict different cyberattacks," in *Data Science and Intelligent Applications*, Singapore, 2021, pp. 43–51.

[24] S. Ali, A. Rauf, N. Islam, H. Farman, and S. Khan, "User profiling: A privacy issue in online public network," *SINDH Univ. Res. J.*, vol. 49, pp. 125–128, Mar. 2017.

[25] A. Sapienza, A. Bessi, S. Damodaran, P. Shakarian, K. Lerman, and E. Ferrara, "Early warnings of cyber threats in online discussions," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 667–674, doi: 10.1109/ICDMW.2017.94.

[26] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *Proc. IEEE Conf. Intell. Secur. Informat. (ISI)*, Sep. 2016, pp. 7–12, doi: 10.1109/ISI.2016.7745435.

[27] M. Bertl, "News analysis for the detection of cyber security issues in digital healthcare a text mining approach to uncover actors, attack methods and technologies for cyber defense," *Young Inf. Scientist*, vol. 4, pp. 1–15, Oct. 2019, doi: 10.25365/yis-2019-4-1.

[28] B. Biswas, A. Mukhopadhyay, S. Bhattacharjee, A. Kumar, and D. Delen, "A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums," *Decis. Support Syst.*, vol. 152, Jan. 2022, Art. no. 113651, doi: 10.1016/j.dss.2021.113651.

[29] B. Song and Y. Suh, "Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety," *J. Loss Prevention Process Industries*, vol. 57, pp. 47–54, Jan. 2019, doi: 10.1016/j.jlp.2018.08.010.

[30] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, pp. 1–12, Jan. 2018, doi: 10.1186/s13673-018-0125-x.

[31] Ö. A. Aslan and R. Samet, "A comprehensive review on malware detection approaches," *IEEE Access*, vol. 8, pp. 6249–6271, 2020, doi: 10.1109/ACCESS.2019.2963724.

[32] A. E. Ibor, F. A. Oladeji, O. B. Okunoye, and O. O. Ekabua, "Conceptualisation of cyberattack prediction with deep learning," *Cybersecurity*, vol. 3, no. 1, pp. 1–14, Jun. 2020, doi: 10.1186/s42400-020-00053-7.

[33] X. Fang, M. Xu, S. Xu, and P. Zhao, "A deep learning framework for predicting cyber attacks rates," *EURASIP J. Inf. Secur.*, vol. 2019, no. 1, May 2019, doi: 10.1186/s13635-019-0090-6.

[34] A. Bilen and A. Özer, "Cyber-attack method and perpetrator prediction using machine learning algorithms," *PeerJ Comput. Sci.*, vol. 7, p. e475, Apr. 2021, doi: 10.7717/peerj-cs.475.

[35] A. Prajapati and S. Gupta, "A survey: Data mining and machine learning methods for cyber security," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 2021, pp. 24–34, Mar. 2021, doi: 10.32628/CSEIT217212.

[36] A. Y. Muaad, G. H. Kumar, J. Hanumanthappa, J. V. B. Benifa, M. N. Mourya, C. Chola, M. Pramodha, and R. Bhairava, "An effective approach for Arabic document classification using machine learning," *Global Transitions Proc.*, vol. 3, no. 1, pp. 267–271, Jun. 2022, doi: 10.1016/j.gltp.2022.03.003.

[37] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, Feb. 2018, doi: 10.1177/0165551516677946.

[38] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[39] B. Gaither, A. Debouse, and C. Huang, "Web page multiclass classification," *SMU Data Sci. Rev.*, vol. 6, no. 1, pp. 1–12, Jun. 2022. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol6/iss1/4

[40] M. Singh, R. Pamula, and S. k. Shekhar, "Email spam classification by support vector machine," in *Proc. Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Sep. 2018, pp. 878–882, doi: 10.1109/GUCON.2018.8674973.

[41] E. Kabir, J. Hu, H. Wang, and G. Zhuo, "A novel statistical technique for intrusion detection systems," *Future Gener. Comput. Syst.*, vol. 79, pp. 303–318, Feb. 2018, doi: 10.1016/j.future.2017.01.029.

[42] M. Elleuch, R. Maalej, and M. Kherallah, "A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition," *Proc. Comput. Sci.*, vol. 80, pp. 1712–1723, Jan. 2016, doi: 10.1016/j.procs.2016.05.512.

[43] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: Review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, Jun. 2019, Art. no. e01802, doi: 10.1016/j.heliyon.2019.e01802.

[44] L. Khairunnahar, M. A. Hasib, R. H. B. Rezanur, M. R. Islam, and M. K. Hosain, "Classification of malignant and benign tissue with logistic regression," *Informat. Med. Unlocked*, vol. 16, 2019, Art. no. 100189, doi: 10.1016/j.imu.2019.100189.

[45] D. J. Wu, T. Feng, M. Naehrig, and K. Lauter. (2015). *Privately Evaluating Decision Trees and Random Forests*. Accessed: Jan. 11, 2023. [Online]. Available: https://eprint.iacr.org/2015/386

[46] M. Abspoel, D. Escudero, and N. Volgushev. (2020). *Secure Training of Decision Trees With Continuous Attributes*. Accessed: Jan. 11, 2023. [Online]. Available: https://eprint.iacr.org/2020/1130

[47] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 10, no. 6, pp. 363–377, Dec. 2017, doi: 10.1002/sam.11348.

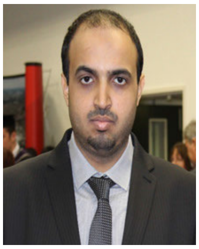
GOHAR MUMTAZ received the B.S. degree in telecommunication engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2013, and the M.S. degree in computer science from the University of Engineering and Technology, Lahore, Pakistan, in 2018. He is currently pursuing the Ph.D. degree in computer science with Superior University, Lahore. From 2014 to 2018, he worked in diverse kind of environment in the domain of network provisioning, maintenance, and operations at well-known organizations, such as, PTCL and Punjab Safe Cities Authority, Lahore. He is currently a Senior Lecturer with Superior University. His research interests include wireless and wired networks and network administration and security. Further, he focuses in 4G and 5G spectrum management and cyber-attacks prediction using machine learning.

SHEERAZ AKRAM received the M.Sc. degree in computer science from the Lahore University of Management Sciences (LUMS), Lahore, Pakistan, and the Ph.D. degree in software engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan. He completed a postdoctoral research training with the University of Pittsburgh, USA, and worked on a project funded through grant U01 HL137159. He completed the STEM Certification and the Certification of Leading People in organization from the University of Pittsburgh. He was with various universities and trained undergraduate and graduate students in industry, academia, and research. He is currently an Associate Professor with the Department of Computer Science, Faculty of Computer Science and Information Technology, Superior University, Lahore. He has published his work in international conferences and journals. His research interests include data science, medical image processing, artificial intelligence in data science, machine learning, deep learning, computer vision, and digital image processing.

MUHAMMAD WASEEM IQBAL received the Ph.D. degree in computer science from Superior University, Lahore, Pakistan. He is currently an Associate Professor with the Software Engineering Department. He is an Active Researcher and has more than 75 research publications in well reputed journals and conferences. Further, he has more than 16 years of teaching and research experience in well-regarded institutions. He specializes in human computer–interaction (HCI), with special interest in adaptive interfaces (AI), user's context, and UX/UI for normal and visual impaired people and user centered design (UCD). His research interests include the Internet of Things (IoT), internet of medical things (IoMT), human centric artificial intelligence (HCAI), semantic relations, and ontological modeling.

M. USMAN ASHRAF received the Ph.D. degree in computer science from King Abdul Aziz University, Saudi Arabia, in 2018. He was a HPC Scientist with the HPC Centre, King Abdul Aziz University. He is currently an Associate Professor and the Head of the Department of Computer Science, GC Women University, Sialkot, Pakistan. His research on exascale computing systems, high performance computing (HPC) systems, parallel computing, and HPC for deep learning and location based services system has appeared in IEEE ACCESS, *IET Software*, *International Journal of Advanced Research in Computer Science*, *International Journal of Advanced Computer Science and Applications*, *International Journal of Information Technology and Computer Science*, *International Journal of Computer Science and Security*, and several International IEEE/ACM/Springer conferences.

KHALID ALI ALMARHABI received the B.Sc. degree in computer science from King Abdul Aziz University, Jeddah, Saudi Arabia, in 2009, the M.Sc. degree in information technology from the Queensland University of Technology, Brisbane, Australia, in 2014, and the Ph.D. degree in computer science from King Abdul Aziz University and the Queensland University of Technology. He is currently an Associate Professor with the Computer Science Department, College of Computing in Al-Qunfudah, Umm Al-Qura University, Saudi Arabia. His research interests include information security, BYODs research, access control policies, information system management, and cloud computing.



AHMED MOHAMMED ALGHAMDI received the B.Sc. degree in computer science and the M.Sc. degree in business administration from King Abdul Aziz University, Jeddah, Saudi Arabia, in 2005 and 2010, respectively, the master's degree in internet computing and network security from Loughborough University, U.K., in 2013, and the Ph.D. degree in computer science from King Abdul Aziz University. He is currently an Assistant Professor with the Software Engineering Department, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia. He has more than 11 years of working experience before attending the academic carrier. His research interests include high-performance computing, big data, distributed systems, programming models, software engineering, and software testing.

ADEL A. BAHADDAD received the B.S. degree in computer science from the Science's College, Saudi Arabia, in 2002, and the M.S. and Ph.D. degrees in information and communication technology from the School of Information and Communication Technology, Griffith University, Brisbane, Australia, in 2012 and 2017, respectively. He is currently an Assistant Professor with the Faculty of Computing and Information Technology, King Abdul Aziz University (KAU), where he has been the Head of the Department of Systems and Educational Programs, Deanship of E-Learning and Distance Education, since 2018. He participated in several executive committees concerned with automating operations with the Educational Curriculum Center and the Strategic Plan of the Strategic Center to achieve the Kingdom's vision with King Abdul Aziz University. His research interests include diffusion and technology adoption and digital transformation, M-service, M-commerce, LMS, and M-governances, and he has many publications in these fields.