Classification and Prediction of Cyber Attacks during COVID-19 using Machine Learning

***Abstract***— The following flare-up of COVID-19 in the world that destroyed the mobility of the whole world also affects cyber security in different ways. During the pandemic, there has been an increase of cyber-attacks due to movement from on-board to work from home all over the seven continents of the world. This research-based paper directs to light up the COVID-19 pandemic with cyber pandemic. We further used data mining for feature extraction from cyber-attacks during pandemic as per the report of Center for Strategic and International Studies (CSIS), then classified using Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). Finally, we conclude this paper with the prediction of which type of cyber-attacks can occur in which part of the world. It is concluded that SVM is a much better classifier than others and it is predicted that Asia is the most affected continent by the cyber-attacks during pandemic.

***Keywords***— Cyber Security, Covid-19, Cyber Attacks, Machine Learning, Data mining, SVM.

1. **Introduction**

The COVID-19 outbreak, which started in China's Hubei province, has spread to a variety of other countries. The World Health Organization (WHO) emergency committee announced a public health emergency on January 30,2020 citing rising case reporting rates in Chinese and foreign locations. The case identification rate fluctuates regularly and can be monitored in almost real time on the Johns Hopkins university website and other forums [1]. China carries the heaviest burden of mortality and morbidity although occurrence in other Asian nations Europe and North America has remained relatively low, that bounds single-stranded large virus that infects many humans along with animals. The symptoms of illness include fever, cough , nasal inflammation, nausea or other respiratory tract infections typically begin after less than a week in symptomatic patients in about 75 percent of patients [2].

Icon

Description automatically generated

**Fig.1. Covid-19 Pandemic with Cyber Pandemic**

The pandemic COVID-19 forced private and public sectors around the globe to move from on-board work to work from home. This movement increased the probability of cyber-attacks [3]. In fact, the cybercrime pandemic is running in the background of COVID-19 pandemic. In April 2021, ransomware attacks increased to 158% as compared with the month of February 2020. Also, more than 500 educational institutes faced this type of attack. As per the report of US Federal Board of Investigation (FBI), cybercrimes increased to 300% and around 6 trillion dollars awaited to be consumed on cybersecurity all over the world till the end of 2021. In fact, IoT devices count will reach around 75 billion till 2025 [4].

In this paper, outfall of cyber-attacks has been predicted based on the data, as per the report of Center for Strategic and International Studies (CSIS) during the pandemic (December 2019 to July 2021). As per the report, we analysed eight types of attacks which occurred in seven continents of the world. We used the features of data mining like n-grams and Bag of Words (BoW) for the feature extraction from the collected data. Machine learning algorithms like Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) are used for data classification [5]. At the end, the paper concludes with the best results of SVM classifiers and Asia (the most affected continent by cyber-attacks) is predicted.

1. **Literature Review**

In December of 2019, COVID-19 was discovered in Wuhan, Hubei Province, China. The surface of coronaviruses is covered in spikes that aid in the attack and binding of living cells. Since this virus shares 86.9% of its DNA with a bat coronavirus, it is thought to have originated in bats. It is a bacterial infection with flu-like symptoms such as a dry cough, sore throat, high fever, and even causes pneumonia, a respiratory illness that can lead to death from respiratory failure. The condition may be mild or serious depending on the circumstances [6]. Coronaviruses have been discovered to infect rodents, rabbits, cats, sheep, goats, turkeys, cattle, and pigs and these animals may transfer to humans in recent studies. If a human meets an infected person, they are at risk of being infected. There are around 177 million confirmed cases and around 3.8 million deaths reported by the World Health Organisation (WHO). Despite extensive attempts, there is no treatment for this disease. The better choices are to get vaccinated, avoidance and management [1] [2].

Covid 19 pandemic introduced cybercrime pandemic. Each sector of the world moved from on board to work from home (online) [4], [7]. Table 1 depicts up to date statistics about internet and social media users during this pandemic covid-19 period. There is an alarming increase of 15.18% of the active social media users.

**Table 1: Internet and Social Media Users Statistics during COVID-19**

| **Stats** | **January 2020** | **July 2021** | **Increase %age** |
| --- | --- | --- | --- |
| **Total Population** | 7.75 billion | 7.87 billion | 1.52 |
| **Unique Mobile Users** | 5.19 billion | 5.27 billion | 1.52 |
| **Internet Users** | 4.54 billion | 4.80 billion | 5.42 |
| **Active Social Media Users** | 3.80 billion | 4.48 billion | 15.18 |
| **Average Internet user time each day** | 6 Hours 43 Minutes | 7 Hours | 5 |

In this situation, covid-19 made inroad for cyber criminals [3], [8]. Some of the facts related to cyber security during pandemic covid 19 are shown in Table 2.

**Table 2: Cyber Security Statistics during COVID-19**

| Cyber Security Statistics COVID-19 | | | |
| --- | --- | --- | --- |
| No. | Fact | Source | Timeline |
| 1 | FBI stated that according to their reports there has been a massive increase of 300% in cyber-crimes, since the beginning of the pandemic caused by Covid-19. | IMC Grupo | May 2020 |
| 2 | The reports and resources ascribed the Covid-19 as the main reason for the 238% rise in cyber-attacks on banks in pandemic as the major cyber-crimes of about 27% are reported from banks and healthcare institutes. | Fintech News | May 2020 |
| 3 | There has been a firm report of data breach or insecurity increment of 58% during the Pandemic in 2020. | Verizon | June 2020 |
| 4 | Pandemic Unemployment Assistance program faced a data security breach of his 33,000 unemployed applicants which caused exposure of data during covid-19 in the month of May 2020. | NBC | May 2020 |
| 5 | Stats shares that there has been a loss of $97.39 to the Americans during the Pandemic which also leads to the cyber-attacks like stimulus check scams. | Atlas VPN | July 2020 |
| 6 | During the Pandemic in April 2020, Google stated that the company obstructs around 18 million of malware and phishing emails on a daily basis which are associated with the coronavirus. | Google | April 2020 |
| 7 | Reports share that the legal and compliance leaders of percentage of about 52% perturbed regarding third party cyber risks because of their remote work during the pandemic of covid-19. | Gartner | April 2020 |
| 8 | There has been an increase of $137,000 in data breaching average cost due to remote work. | IBM | August 2020 |
| 9 | During the covid-19 pandemic an estimated stat share that there are 47% of the adduce employees who fall prey to the scam of phishing while they are working from their homes. | Tessian | July 2020 |
| 10 | In covid-19 there has been a massive amount of 81% of cyber-security professionals that frequently report about the changing of functions in their job. | ISC | April 2020 |
| 11 | In April 2020, Zoom company claimed that five hundred thousand accounts of their users had been compromised during the pandemic and sold to a forum on dark web. | CPO Magazine | April 2020 |
| 12 | In the crises of covid-19 a report appears which shows that there has been a humongous increase of about 630% in cloud based cyber-attacks during the time period of January to April in 2020. | Fintech News | April 2020 |
| 13 | Around 20% of the security breaches in organizations are caused by the remote workers. | Malwarebytes | August 2020 |
| 14 | During the pandemic of covid-19, there has been an increase of 600% in cyber-crimes because of the global pandemic that covers and affects the whole world. | Purple Sec | May 2020 |

In this paper, we predicted outbreak cyber-attacks during COVID-19. We used data mining and machine learning algorithms to obtain optimized results. The authors in the [9] predicted the outbreak of COVID-19 using machine learning. In this paper, the author worked on real time forecasting of COVID-19 in one province of China. In [10], [11], Deep learning methodology is used for forecasting cyber-attacks based on the captured data from network traffic. In another paper [12], cyber-attacks methods and committers have been predicted using Support Vector Machine, a machine learning algorithm. In [13], [14], authors suggested different machine learning approaches like Bayesian network, Decision Tree, Clustering, Artificial Neural Networks (ANN) in cyber security to detect cyber threats. In another paper [15], the authors worked on a dataset, collected from five different enterprises in Korea. They also used data mining and machine learning algorithms, like n-gram and bag of words for feature engineering and Bayesian network for classifiers.

Data Mining and text analysis are very well-known techniques being used to detect vulnerability in cyber security and that is the reason we used n-gram and BoW for feature extraction, in our model. For the classifier, we used NB, SVM, LR and RF.

1. **Methodology**

The main concept in our research work is the centralised classifier, as shown in Fig.2. used to collect data from the seven continents of the world. The data sets are used to train this classifier and eventually achieve better performance rather than using separate classifiers for each continent.

Text

Description automatically generated

**Fig.2. Centralised Classifier for data collection from 7 continents**

* 1. **Dataset interpretation**

The Dataset is a type of cyber-attacks, occurring in 7 continents of the word during COVID-19 (December 2019 to July 2021), as per the report of Center for Strategic and International Studies (CSIS). Datasets are classified continent wise and there are a total 396 attacks. The distribution of the dataset is shown in Table 3. Also, the graphical representation for continent wise number of cyber-attacks is shown in Fig.3. Each attack entry (in a single row) has the following details. From the dataset, we chose the attack type as labelled data.

* Year
* Month
* Name of the Continent
* Name of the Country
* Type of cyber-attack

**Table 3: Distribution of the Data**

| **Continent** | **Count of Attacks** | **Type of Attack** | **Count of Type of Attack** |
| --- | --- | --- | --- |
| **Africa** | 7 | **APT** | 99 |
| **Asia** | 170 | **DDoS** | 20 |
| **Europe** | 142 | **Espionage** | 13 |
| **North America** | 17 | **Malware** | 85 |
| **Oceania** | 8 | **Man-in-Middle** | 3 |
| **South America** | 3 | **Phishing** | 170 |
| **United States** | 49 | **SQL Injection** | 1 |
|  |  | **Zero-Day Exploit** | 5 |
| **Total** | 396 | **Total** | 396 |

Chart, bar chart, histogram

Description automatically generated

**Fig.3. Continent wise number of cyber-attacks**

* 1. **Research Problem**

In this paper, we identified, investigated, and solved how to forecast the name of the continent/country based on the type of cyber-attack. For this, we used 4 types of classifiers to check efficiency of the classifiers for the collected dataset.

* 1. **Method of Analysis and Classification**

This section is further divided into three sub-sections.

1. **Data collection and preprocessing**

The data is collected from the report of Center for Strategic and International Studies (CSIS). We sorted out ambiguities due to null values in the dataset. Then we chose two columns (Continent and Country) with labelled data (type of cyber-attack) to solve our research problem. At this stage, our dataset is ready for feature extraction.

1. **Feature extraction**

For feature extraction, the most common features of data mining are n-gram and BoW. We used the concept of unigram and bi-gram models. For uni-gram, n=1. For example, “phishing” is a word, and all the attacks will be extracted containing this word from the dataset. We used Term Frequency – Inverse Document Frequency (TF-IDF), a technique based on the BoW model used for text vectorization (in our case feature vector (396,32)).

1. **Classification**

We used four different classifiers of Machine learning.

* 1. Naïve Bayes (NB),
  2. Support Vector Machine (SVM),
  3. Logistic Regression (LR),
  4. Random Forest (RF),

1. **Experiments and results**

We used four different types of classifiers to perform experimental study to explore the ability of classifiers. The output of the classifier is to predict the name of the continent/country based on the type of cyber-attack. 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, first we trained our model through 396 attacks and then tested and predicted the output based on the type of cyber-attack. Firstly, we used an SVM classifier to predict the output. The evaluated results are shown in Table 4, while the confusion matrix for the SVM model is shown in Fig. 4. The Accuracy of this model is 0.67 and the values of Precision, Recall and F1 measures against Africa, South America and the United States are all zero (0.00) because of the very limited number of cyber-attacks in these continents.

**Table 4: Model evaluated using SVM**

|  | **SVM** | | |
| --- | --- | --- | --- |
| **Accuracy** | 0.67 | | |
| **Class** | **Precision** | **Recall** | **F1-measure** |
| **Africa** | 0.00 | 0.00 | 0.00 |
| **Asia** | 0.97 | 0.65 | 0.78 |
| **Europe** | 0.53 | 0.98 | 0.69 |
| **North America** | 0.83 | 0.83 | 0.83 |
| **Oceania** | 1.00 | 0.50 | 0.67 |
| **South America** | 0.00 | 0.00 | 0.00 |
| **United States** | 0.00 | 0.00 | 0.00 |

Graphical user interface, application

Description automatically generated

**Fig.4. SVM Confusion Matrix**

Secondly, we used the NB Classifier to check its efficiency against our model. The evaluated results 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 less than SVM and the values of precision, recall and f1-measure are zero (0.00) against four continents.

**Table 5: Model evaluated using NB**

|  | **NB** | | |
| --- | --- | --- | --- |
| **Accuracy** | 0.56 | | |
| **Class** | **Precision** | **Recall** | **F1-measure** |
| **Africa** | 0.00 | 0.00 | 0.00 |
| **Asia** | 0.57 | 0.98 | 0.72 |
| **Europe** | 0.95 | 0.47 | 0.63 |
| **North America** | 1.00 | 0.17 | 0.29 |
| **Oceania** | 0.00 | 0.00 | 0.00 |
| **South America** | 0.00 | 0.00 | 0.00 |
| **United States** | 0.00 | 0.00 | 0.00 |

Graphical user interface, text, application, Teams

Description automatically generated

**Fig.6. NB Confusion Matrix**

Thirdly, we used the LR classifier, and the results are quite like the SVM model. But the precision value against Asia is 0.95 while it is 0.97 when we used SVM. Detailed evaluated result is depicted in Table 6 and the confusion matrix in Fig.7.

**Table 6: Model evaluated using LR**

|  | **LR** | | |
| --- | --- | --- | --- |
| **Accuracy** | 0.67 | | |
| **Class** | **Precision** | **Recall** | **F1-measure** |
| **Africa** | 0.00 | 0.00 | 0.00 |
| **Asia** | 0.95 | 0.65 | 0.77 |
| **Europe** | 0.53 | 0.98 | 0.69 |
| **North America** | 1.00 | 0.50 | 0.67 |
| **Oceania** | 1.00 | 0.50 | 0.67 |
| **South America** | 0.00 | 0.00 | 0.00 |
| **United States** | 0.00 | 0.00 | 0.00 |

Graphical user interface, application

Description automatically generated

**Fig.7. LR Confusion Matrix**

Finally, and fourthly, we used the RF classifier and the evaluated results with the accuracy of 0.51 are shown in Table 7. The confusion matrix is shown in Fig.8.

In the end, we comparatively concluded that the SVM classifier is better than other classifiers. The Accuracy measure for all the four classifiers is shown in Fig.9.

**Table 7: Model evaluated using RF**

|  | **RF** | | |
| --- | --- | --- | --- |
| **Accuracy** | 0.51 | | |
| **Class** | **Precision** | **Recall** | **F1-measure** |
| **Africa** | 0.00 | 0.00 | 0.00 |
| **Asia** | 0.97 | 0.65 | 0.78 |
| **Europe** | 0.53 | 0.98 | 0.69 |
| **North America** | 0.83 | 0.83 | 0.83 |
| **Oceania** | 1.00 | 0.50 | 0.67 |
| **South America** | 0.00 | 0.00 | 0.00 |
| **United States** | 0.00 | 0.00 | 0.00 |

Graphical user interface, application

Description automatically generated

**Fig.8. RF Confusion Matrix**

Chart, box and whisker chart

Description automatically generated

**Fig.9. Accuracy measure for different classifiers**

1. **Conclusion**

In this paper, we focussed on the research based on cyber-attacks during COVID-19 (December 2019 to July 2021). We collected data from the report of Center for Strategic and International Studies (CSIS). We analysed and classified the data using data mining and machine learning algorithms. We used four different classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) and predicted the output (name of the continent based on type of cyber-attacks). We also predicted which continent is more affected by the cyber-attacks during COVID-19. It is concluded that SVM is better than other classifiers against our model and Asia is the most affected continent by cyber-attacks during pandemics.

1. **References**

[1] T. P. Velavan and C. G. Meyer, “The COVID-19 epidemic,” *Trop Med Int Health*, vol. 25, no. 3, pp. 278–280, Mar. 2020, doi: 10.1111/tmi.13383.

[2] Prof. I. Ali and O. Alharbi, “COVID-19: Disease, management, treatment, and social impact,” *Science of The Total Environment*, vol. 728, p. 138861, Apr. 2020, doi: 10.1016/j.scitotenv.2020.138861.

[3] H. Lallie *et al.*, *Cyber Security in the Age of COVID-19: A Timeline and Analysis of Cyber-Crime and Cyber-Attacks during the Pandemic*. 2020.

[4] H. Hejase, H. Kazan, A. Hejase, and I. Moukadem, “Hejase et al. Cyber Security paper,” *Computer and Information Science*, vol. Vol. 14, pp. 10–25, Mar. 2021, doi: 10.5539/cis.v14n2p10.

[5] S. F. Ardabili *et al.*, “COVID-19 Outbreak Prediction with Machine Learning,” *Algorithms*, vol. 13, no. 10, 2020, doi: 10.3390/a13100249.

[6] T. Fisayo and S. Tsukagoshi, “Three waves of the COVID-19 pandemic,” *Postgrad Med J*, vol. 97, no. 1147, p. 332, May 2021, doi: 10.1136/postgradmedj-2020-138564.

[7] L. Manikam *et al.*, “Online community engagement in response to COVID-19 pandemic,” *Health Expectations*, vol. 24, no. 2, pp. 728–730, Apr. 2021, doi: 10.1111/hex.13194.

[8] M. Baz *et al.*, “Impact of COVID-19 Pandemic: A Cybersecurity Perspective,” *Intelligent Automation and Soft Computing*, vol. 27, p. 641, Mar. 2021, doi: 10.32604/iasc.2021.015845.

[9] D. Liu *et al.*, *A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models*. 2020.

[10] A. Ibor, F. Oladeji, O. Okunoye, and O. Ekabua, “Conceptualisation of Cyberattack prediction with deep learning,” *Cybersecurity*, vol. 3, pp. 1–14, Jun. 2020, doi: 10.1186/s42400-020-00053-7.

[11] X. Fang, M. Xu, S. Xu, and P. Zhao, “A deep learning framework for predicting cyber attacks rates,” *EURASIP Journal on Information Security*, vol. 2019, May 2019, doi: 10.1186/s13635-019-0090-6.

[12] A. Bilen and A. Özer, “Cyber-attack method and perpetrator prediction using machine learning algorithms,” *PeerJ Computer Science*, vol. 7, p. e475, Apr. 2021, doi: 10.7717/peerj-cs.475.

[13] R. Das and T. H. Morris, “Machine Learning and Cyber Security,” in *2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, Dec. 2017, pp. 1–7. doi: 10.1109/ICCECE.2017.8526232.

[14] N. Parati and P. Anand, “Machine Learning in Cyber Defence,” *International Journal of Computer Sciences and Engineering*, vol. 5, pp. 317–322, Dec. 2017, doi: 10.26438/ijcse/v5i12.317322.

[15] A. Mohasseb, B. Aziz, J. Jung, and J. Lee, *Predicting CyberSecurity Incidents Using Machine Learning Algorithms: A Case Study of Korean SMEs*. 2019. doi: 10.5220/0007309302300237.