Code Text-based Virus and Non-virus Classification and Comparative Analysis of Machine Learning Algorithms

Anchit Akshaansh[1], Vipin Kumar[2] and Aaryan Kumar Hrithik[3]

Address: *[1, 2,3]Department of computer science and information Technology, Mahatma Gandhi Central University, Bihar-845401, India*
[1]Email: anchit.1650@gmail.com [2]Email: rt.vipink@gmail.com [3]Email: aaryanhrithik7@gmail.com

Abstract

In today's world, The Antivirus is capable of handling all the threats and vulnerabilities that are fed into the software after getting that knowing vulnerability, but every time the attack. To counter new threat features, Processing for making a new technology, where the data set is analysed the type of Viruses, Malware, Trojan Horses, Backdoors, Ransomware, and Rootkits used for attacking the victim. The purpose of this project is to make a new type of software (Anti-Virus) that will be capable of tracing all the threats where the aim is to analyse every code that is previously available or available in the market, or that is not in the software. In this research, a novel dataset of virus and non-virus has been created i.e., 97 codes of applications. The text of the codes is used to extract the features text pre-processing techniques and 26 different algorithms are utilized for the extensive analysis of classification over four parameters i.e., accuracy, balance accuracy, AUC-ROC, and F1-Score. The highest classification accuracy has been achieved by the four classifiers equally (i.e., 78.95%) which are AdaBoost, Decision tree, GaussianNB and Bagging.

Keywords: Cybersecurity, Virus, Malware, Trojan horse, Malicious, Backdoor, Ransomware, Threat, Text-Mining, Classification, Machine Learning.

1. INTRODUCTION

As the world is moving towards Digital called Cyberspace. Cyberspace is where work done digitally, like using Social Media Apps, Em@ils, Net Banking, Work from Home, etc., is done. Everything done in cyberspace is stored and has the vulnerability threat of being exploited by the Attackers (Bad Hackers). To secure digital devices, there is much software like Antivirus. Still, not every technology is efficient for all threats as every minute in the world; there are Approximately 2 lakhs to 3 lakhs or more attacks practiced by hackers. You can check this on the online threat map where attacking types and numbers are shown which are updated every second. Therefore, to cope with new technologies of attack, new security technologies are developed its counter. However, it is not the only method to counter that type of attack in industries of the world.

There are several types of security available in cyberspace like cybersecurity, information security, network security, malware, etc. The goal of cybersecurity is to protect from the risks or to save from unauthorized access or alteration. The main source & channel is the internet where people do business. The attack is aimed at assessing, altering, and removing sensitive data, extracting money, or interfering with normal business operations [1]. The aim of information security is to protect the privacy of information and hardware that manages saves and transmits that information [2]. The user got attached by malware i.e., Breach of the network through vulnerability, clicking on a suspicious email connection, or installing risky software. Once entering the network, it receives sensitive data; due to flaws in the framework, more harmful software can be made. Blocking access to strong business networks [3]. Network security is to protect users on the network, when the network achieves this, the potential threat gets blocked from the introduction. It contains a firewall that blocks unauthorized access to a network for secure remote access. Snort is one of the Network Security which includes IP tracking, Tracing, and Intrusion Detection mechanisms [4].

The code of virus and non-virus are very important content to understanding the type of application. Therefore, this research focuses to classify the virus and non-virus applications based on their codes. The author has collected a novel dataset of virus and non-virus code and extracted the features using text preprocessing techniques. The twenty-six (26) machine learning algorithms are used to do the extensive analysis for the classification task, where accuracy, balance accuracy, AUC-ROC, and F1-Score are measured for the comparison of distinct classifiers. It has been observed that four classifiers (i.e., Adaboost, Decision Tree, GaussianNB, and Bagging).

The novelty of the research work is as follows:

- The author has collected a novel dataset of the 97 codes in the form of the text of virus (50) and non-virus (47) applications;
- The features from the text related to virus and non-virus have been identified manually after pre-processing of instructions of code because pre-processing of the code is different from the normal text pre-processing. Then the total number of features got extracted from the codes is 637 including virus and non-virus applications;
- Extensive comparative analysis of machine learning algorithms (26 algorithms) has been done with three performance measures i.e., AUC, ROC, and F1-Score;
- The various future research scope has been identified based on text classification of virus and non-virus application;

In this research paper: Section I has the introduction of cybersecurity, applications, short description of proposed methodology and novelty of the work. In next section II has the brief description of machine leaning and their related application in cybersecurity along with brief description of various machine-learning algorithms. Section III describe the literature review related the work. Methodology of proposed word with flowchart has been discussed in the Section IV and Section V has the results and their analysis over various parameter of the performances. Lastly, Section VI has the conclusion of research work.

2. LITERATURE REVIEW

Attackers are skilled, circumvent security measures and allow them to go undetected for extended periods. Worse, attack mechanisms are becoming commoditized, making them easier to distribute without requiring a deep understanding of how to create. The project on which work is going on is “Cybersecurity-Text Mining Using Lazy Predict (Classification): Machine Learner”, where I must take a Dataset of text files that contain Viruses and Non-Viruses [16, 17]. Dataset will be used for analyzing either Virus or Non-virus text files by applying Machine Learning Python Code for feature extraction, Classifier, F1 Score, Recall, Precision, Accuracy, Confusion Metric, Heat Map of each Classifier (KNN—Nearest Neighbours Classifiers [18], SVM-Support Vector Machine [19], DT-Decision Tree

[20], MLP-Multilayer-Perceptron Classifier [21], Gaussian NB-Gaussian Naive Bias) [22]. Creating the new dataset is not easy. When I was creating a dataset for my Machine Learning Model, I got some difficulties searching the virus and Non-Virus Code [23]. The main problem is saving it in windows as Now in these days Windows does not allow to save of virus files. If you forcefully save it in windows defender will detect and delete that file forcefully. Therefore, to solve that problem, I have saved in zipped on Google Drive and GitHub. The main thing here is to make a dataset.

3. METHODOLOGY

Figure 1 shows the flowchart of the proposed work for virus and non-virus classification tasks.

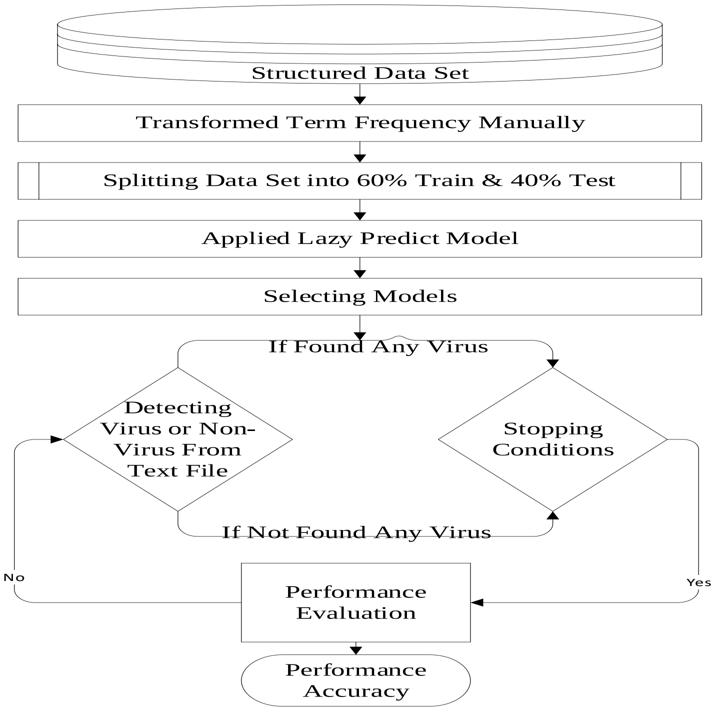
Structured Data Set
Transformed Term Frequency Manually
Splitting Data Set into 60% Train & 40% Test
Applied Lazy Predict Model
Selecting Models
If Found Any Virus
Detecting Virus or Non-Virus From Text File
Stopping Conditions
If Not Found Any Virus
No
Yes
Performance Evaluation
Performance Accuracy

Fig. 1: Flowchart of proposed work for virus and non-virus classification.

The virus and non-virus text classification task contain the following subtasks the collection of data, preprocessing of the text of code, data set sampling for the training and testing, deployment of different machine learning algorithms along with hyperparameter tuning using validation process, and final evaluation of performance using distinct measures.

4. EXPERIMENTS

4.1. *Collection of datasets:*

Dataset collection is novel & obtained from different sources using Tor Browser (A particular browser for ethical hackers to gain any information about hacking and new Virus, Malware, Ransomware, Trojan, etc. The list of the virus and non-virus are shown in Table 1. Here is the list of categories that have been used to obtain the different virus and non-virus: demands without considering the data volume. Taking into consideration this ratio, Green IT Technologies have important benefits in terms of:

- ***Virus:*** *A software that causes damage to data and software [24];*

No.	Virus	Non-Virus	No.	Virus	Non-Virus
1.	Adaptor Info	AccessinArrays1	26.	Matrix	ListDirectoriesUsingCmd
2.	Application Bomber	AccessinArrays2	27.	MIS17Port445	LocalNGlobalVariables
3.	Attempt PWN	Arithmetic Operators	28.	Non-WorkingTXTFiles	LoggingErrorMessagesToAnotherFile
4.	Bugs& Ransomware	Block_usb	29.	PC_Virus.C	LoopCMDLineArguments
5.	Computer Crash	CommandLinePrinter Control	30.	PCCrashes	MAS_1.4_AIO_CRC32_9A7B5B05{CrackHAsh}
6.	ComputerShutsDownWhenTurnedOn	CreatingAlias	31.	PcCrashForever	ModifyingArray
7.	CryptAcquireContent	CreatingAlias1	32.	PcShutDown	ModifyingExistingArray
8.	DriveContent Delete	Creating Array	33.	PoppingCD	NoFormComments
9.	Endless Enter	CreatingArrayStructure	34.	ProcessCreator	NumericValuesSet ASwitch
10.	Endless Notepad	CurrentDirectoryWithPrompts&Warnings	35.	RegistryDeleter	Office
11.	Eternal Blue MIS17 Ransomware	Date	36.	SEO	Patch
12.	Eternal Blue Worm	DateFormatYearMonthDay	37.	ShellSH	Pause
13.	Fake Windows Error With Notepad	Debbuging	38.	ShutDownComputer EveryTimeStart	PrinterCMD
14.	FolderFoolder	Deleting Alias	39.	ShutInternet Permanently	RemCMD
15.	EternalBlueMS17Ransomware	EchoCMD	40.	System32Delete	ReplacingAlias
16.	EternalBlue Worm	ErrorLevel	41.	SimpleHarmlessVirus	RunningProcessLists
17.	FormatDrives	ErrorLevelToDetectErrorNLog	42.	TextToAudio Convertion	SetupComplete
18.	Goliate Hidden Tear Ransomware	FindComputers&Logged Users	43.	ToggleButton	StartingNewprocess
19.	Green001-Something Ransomware	FunctionDefination	44.	UserAccountFlooder	TestPrinterExistence
20.	HarmlessCDRom Virus	FunctionDefination1	45.	VIRUS-VBS CODE	Unblock_usb
21.	ILoveYou	HelloCMD	46.	VirusBasicFormat	UsingTheStatement
22.	InternetDisabler	IteratingArray	47.	VirusToTestAntivirus	ViewrunningProcessList
23.	InternetOpenTypes Direct	JavaEnvironment Variables	48.	WindowsCrash	
24.	IP	KillingParticularProcess	49.	WindowsHacker	

25.	IPScanLoop	LengthofArray	50.	WindowsLogOff

Table 1:List of Virus & Non-Virus Dataset

4.2. Pre-processing of text data of the code:

Dataset set is transformed into the numerical form using the tf-idf formula [28]. The first dataset is converted into term frequency for the further conversion of term frequency-inverse document frequency, where 637 number of features are extracted for the 97 samples for the binary classification task. The tokens have identified manually because for the coding there are no tokenization tools are available to the best of my knowledge. After pre-processing, a structure labelled dataset has been prepared with 97 rows and 637 columns before deploying the machine learning algorithms.

4.3. Sampling of the dataset before deploying the machine learning algorithms:

The labelled structured dataset has been splitted into 60% and 40% samples for training and testing purposes. It has been resampled while all iterations of the experiments.

4.4. Deployment of machine learning algorithms:

These steps of the framework, deploy the machine learning algorithms and tune hyperparameter while training the algorithms, then predict the test data to get the performances of the classifiers corresponding to the given measure(s). In this research 26 machine learning algorithms are utilized to analyse the performance i.e., AdaBoost, Decision Tree, Naïve bayes, Gaussian NB, Bagging, BernoulliNB, XGB, SVC, NuSVC, SGD, Ridge CV, Ridge, Random Forest, Quadratic Discriminant analysis, perceptron, Passive aggressive, Linear SVC, Nearest Centroid, Logistic Regression, LDA, Label Spreading, Label Propagation, KNN, Extra Tree, and Calibrated CV.

4.5. Evolution of machine learning algorithms:

The performance of the algorithms may evaluate with different parameters like overall accuracy, balance accuracy, precision, recall, F1-Score, etc. This may utilize a combination of measures to evaluate the obtained classifiers that has obtain while training. The analysis of the classifiers for the classification task is performed based on Accuracy, balance accuracy, AUC-ROC, and F1-Score.

5. RESULT AND ANALYSIS

5.1. Description of Results:

Fig. 2 show the accuracies and balance accuracies comparison of various machine learning algorithms and the ROC-AUC and F1-Score based performances of the classifiers, the x- axis and y-axis are denoted as list of classifiers and performance measures, respectively.

5.2. Analysis of Results:

5.2.1 Accuracy:

There are four classifiers of Lazy Predict, which have the highest equal accuracy of 78.95% on both virus and non-virus datasets AdaBoost Classifier, Decision Tree Classifier, Gaussian NB, and Bagging Classifier. While the second highest best-performing classifiers are SGD Classifier, Perceptron, Nearest Centroid, Label Spreading, and Label Propagation, with an accuracy of 68.42%. The third highest performing classifier is BernoulliNB, with an accuracy of 43.36%. While the remaining classifiers are performing the same with the lowest classification accuracy of 31.57%.

5.2.2 Balanced Accuracy:

There are four classifiers of Lazy Predict, which have the highest equal Balanced Accuracy of 66.67% on both virus and non-virus datasets AdaBoost Classifier, Decision Tree Classifier, Gaussian NB, and Bagging Classifier. While the second highest best performing classifier is BernoulliNB, with an accuracy of 57.05%. While the remaining classifiers are performing the same, with the lowest classification Balanced Accuracy of 50%.

5.2.3 ROC-AUC:

The result of ROC is the same as the Balanced Accuracy, i.e., there are four classifiers of Lazy Predict, which have the highest equal Balanced Accuracy of 66.67% on both virus, and non-virus datasets are AdaBoost Classifier, Decision Tree Classifier, Gaussian NB, and Bagging Classifier. While the second highest best performing classifier is BernoulliNB, with an accuracy of 57.05%. While the remaining classifiers are performing the same, with the lowest classification Balanced Accuracy of 50%.

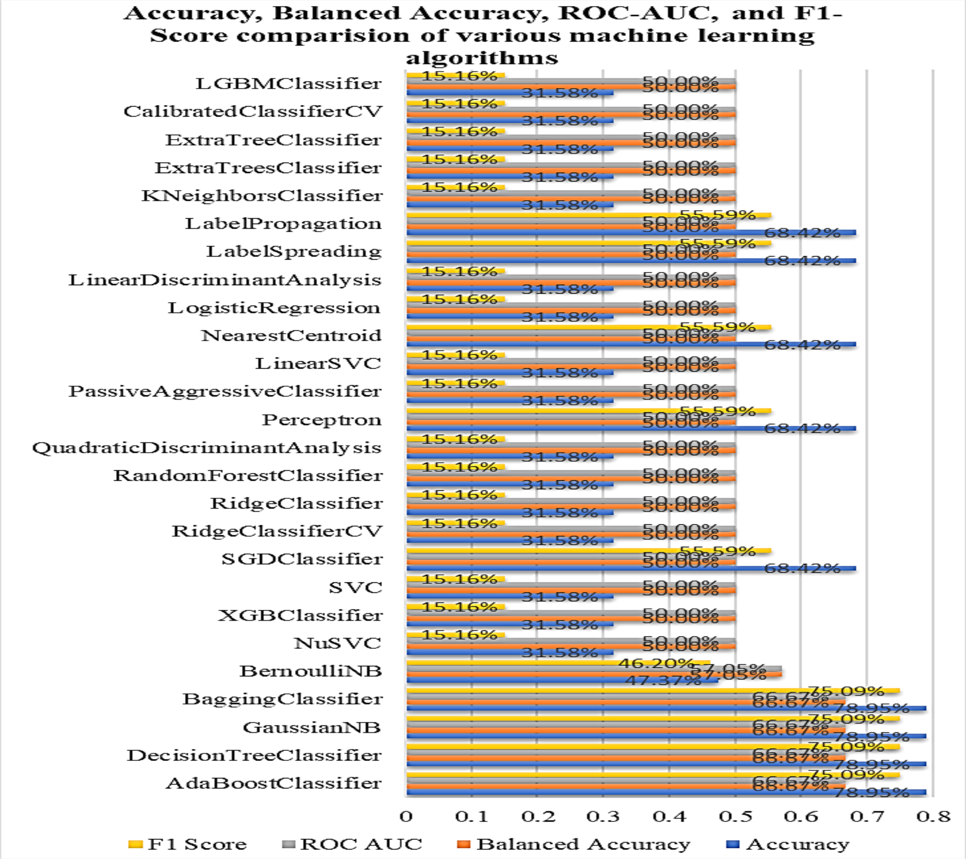

Fig. 2: *Comparison of Various Machine Learning Algorithms*

5.2.4 F1-Score:

Here, AdaBoost Classifier, Decision Tree Classifier, Gaussian NB, and Bagging Classifier are the best performers, with an F1-Score of 75.09%. While the second highest best-performing classifiers are SGD Classifier, Perceptron, Nearest Centroid, Label Spreading, and Label Propagation, with an accuracy of 55.60%. The third highest performing classifier is BernoulliNB, with an accuracy of 46.20%. While the remaining classifiers are performing the same with the lowest classification accuracy of 15.15%.

Conclusion of the above analysis show that AdaBoost, Decision Tree, Gaussian NB, and Bagging Classifier are the best performer among all four measures accuracy, balance accuracy, ROC-AUC, and F1-Score. These classifiers may be utilised for the text classification of virus and non-virus code directly.

6. CONCLUSION

There are four classifiers which have the highest Accuracy of 78.95%, Balanced Accuracy of 66.67%, ROC of 66.67, and F1-Score of 75.09% on both virus and non-virus datasets. According to the result of Lazypredict Classifiers, which have the highest equal accuracy of 78.95% on both virus and non-virus datasets are AdaBoost Classifier, Decision Tree Classifier, GaussianNB, and Bagging Classifier, indicates that these classifiers are efficient in detecting the virus from the trained & test dataset. While the second highest best-performing classifiers are SGD Classifier, Perceptron, Nearest Centroid, Label Spreading, and Label Propagation with an accuracy of 68.42% . The third highest performing classifier is BernoulliNB, with an accuracy of 43.36%, indicating that it has an efficiency of detection chance of 43.36%, in contrast with the above two highest & second highest accuracy. While the remaining classifiers are performing the same with the lowest classification accuracy of 31.57%, indicating that they can detect. Still, viruses may be detectable or not according to a given accuracy.

7. REFERENCES

[1] D. Craigen, N. Diakun-Thibault, and R. Purse, “Defining cybersecurity,” *Technology Innovation Management Review*, vol. 4, no. 10, 2014.

[2] R. Anderson and T. Moore, “The economics of information security,” *Science (1979)*, vol. 314, no. 5799, pp. 610–613, 2006.

[3] N. Idika and A. P. Mathur, “A survey of malware detection techniques,” *Purdue University*, vol. 48, no. 2, pp. 32–46, 2007.

[4] M. Kaeo, *Designing network security*. Cisco Press, 2004.

[5] A. Handa, A. Sharma, and S. K. Shukla, “Machine learning in cybersecurity: A review,” *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 9, no. 4, p. e1306, 2019.

[6] A. Handa, A. Sharma, and S. K. Shukla, “Machine learning in cybersecurity: A review,” *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 9, no. 4, p. e1306, 2019.

[7] J.-Y. Wu, “Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement,” *Comput Educ*, vol. 163, p. 104066, 2021.

[8] I. F. Kilincer, F. Ertam, and A. Sengur, “Machine learning methods for cyber security intrusion detection: Datasets and comparative study,” *Computer Networks*, vol. 188, p. 107840, 2021.

[9] S. Ismail and H. Reza, “Evaluation of Na\"\ive Bayesian Algorithms for Cyber-Attacks Detection in Wireless Sensor Networks,” in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 283–289.

[10] G. Wang, D. Tse, Y. Cui, and H. Jiang, “An Exploratory Study on Sustaining Cyber Security Protection through SETA Implementation,” *Sustainability*, vol. 14, no. 14, p. 8319, 2022.

[11] S. Ismail and H. Reza, “Evaluation of Na\"\ive Bayesian Algorithms for Cyber-Attacks Detection in Wireless Sensor Networks,” in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 283–289.

[12] H. Berry, M. A. Abdel-Malek, and A. S. Ibrahim, “A Machine Learning Approach for Combating Cyber Attacks in Self-Driving Vehicles,” in *SoutheastCon 2021*, 2021, pp. 1–3.

[13] M. Choubisa, R. Doshi, N. Khatri, and K. K. Hiran, “A Simple and Robust Approach of Random Forest for Intrusion Detection System in Cyber Security,” in *2022 International Conference on IoT and Blockchain Technology (ICIBT)*, 2022, pp. 1–5.

[14] S. S. A. Krishnan, A. N. Sabu, P. P. Sajan, and A. L. Sreedeep, “SQL Injection Detection Using Machine Learning,” *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS*, vol. 11, no. 3, pp. 300–310, 2021.

[15] S. H. Majidi, S. Hadayeghparast, and H. Karimipour, “FDI attack detection using extra trees algorithm and deep learning algorithm-autoencoder in smart grid,” *International Journal of Critical Infrastructure Protection*, vol. 37, p. 100508, 2022.

[16] C. Florackis, C. Louca, R. Michaely, and M. Weber, “Cybersecurity risk,” 2020.

[17] S. Skryl *et al.*, “Assessing the response timeliness to threats as an important element of cybersecurity: Theoretical foundations and research model,” in *Conference on Creativity in Intelligent Technologies and Data Science*, 2019, pp. 258–269.

[18] P. P. Kaur and S. Singh, “Classification of Herbal Plant and Comparative Analysis of SVM and KNN Classifier Models on the Leaf Features Using Machine Learning,” pp. 227–239, 2021, doi: 10.1007/978-981-16-1048-6_17.

[19] E. Burnaev and D. Smolyakov, “One-class SVM with privileged information and its application to malware detection,” in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 273–280.

[20] J. L. Leevy, J. Hancock, R. Zuech, and T. M. Khoshgoftaar, “Detecting cybersecurity attacks across different network features and learners,” *J Big Data*, vol. 8, no. 1, pp. 1–29, 2021.

[21] T. T. Teoh, G. Chiew, E. J. Franco, P. C. Ng, M. P. Benjamin, and Y. J. Goh, “Anomaly detection in cyber security attacks on networks using MLP deep learning,” in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 2018, pp. 1–5.

[22] M. Yousefi-Azar, V. Varadharajan, L. Hamey, and U. Tupakula, “Autoencoder-based feature learning for cyber security applications,” in *2017 International joint conference on neural networks (IJCNN)*, 2017, pp. 3854–3861.

[23] A. Mitra, Digital security: cyber terror and cyber security. Infobase Publishing, 2010.

[24] P. Szor, The Art of Computer Virus Research and Defense: ART COMP VIRUS RES DEFENSE _p1. Pearson Education, 2005.

[25] M. T. Signes-Pont, A. Cortés-Castillo, H. Mora-Mora, and J. Szymanski, “Modelling the malware propagation in mobile computer devices,” *Comput Secur*, vol. 79, pp. 80–93, 2018.

[26] T. B. Slayton, “Ransomware: The virus attacking the healthcare industry,” *Journal of Legal Medicine*, vol. 38, no. 2, pp. 287–311, 2018.

[27] A. Riyana, B. Santoso, and R. Hartono, “Trojan malware analysis using reverse engineering method in Windows 7,” *Technium Soc. Sci. J.*, vol. 30, p. 775, 2022.

[28] R. Rawat, V. Mahor, S. Chirgaiya, R. N. Shaw, and A. Ghosh, “Analysis of darknet traffic for criminal activities detection using TF-IDF and light gradient boosted machine learning algorithm,” in *Innovations in electrical and electronic engineering*, Springer, 2021, pp. 671–681.

Biographies

Anchit Akshaansh is currently working as Assistant Professor in the Department of Computer Science and Engineering, Meerut Institute of Technology, Meerut, UP India. Author has completed his B.Tech (2017) in Computer Science and Engineering form Biju Patnaik University of Technology, Rourkela, Odisha and Received M.Tech (2022) in Computer Science and Engineering from Mahatma Gandhi Central University,Motihari Bihar. He is focused in CyberSecurity, Machine Learning, Deep Learning Text Mining, Image Processing etc. Author Aim is to Develop Machine and Deep Learning Techniques for Cybersecurity and has in field of Image Processing. Author is looking further for PhD in his interest field.

Dr. Vipin Kumar is currently working as an Assistant Professor in the Department of Computer Science and Information Technology, Mahatma Gandhi Central University, Bihar India. Author has Completed B.Tech (2009) in Computer Science and Engineering from National Institute of Technology Allahabad (NITA), UP, India. In addition, he has received M.Tech (2012) and PhD (2015) in Computer Science from JNU.New Delhi, India. Teaching and Research areas are Machine Learning, Data Science, Data Mining, Deep Learning and Big Analytics. Seven (7) Years of Teaching Experience for UG, PG and PhD. He teaches Machine Learning, Data Mining, Data Science Soft Computing, Text Mining, Sentiment Analysis, Data Base Management Systems, Data Structure, Python Programming, C Programming, Combinatorial Optimization, etc. His associated and reputed journals are Information Science, Information Fusion, etc. and review several research papers. He is focused on research to Develop Machine Learning Techniques for a High-Dimensional Dataset for Classification Tasks, Image processing Using Machine Learning and Deep Learning for Seeds & Leaf Detection in Agriculture Fields and Semi-Supervised Learning Tasks for Remote Sensing, Developing Algorithms for Multi-View Learning for the Classification Task.

Aaryan Kumar Hrithik has completed his B.Tech (2019) in Computer Science and Engineering from I. K Gujral Punjab Technical University, Jalandhar and Received M.Tech (2022) in Computer Science and Engineering from Mahatma Gandhi Central University, Motihari Bihar. He is focused in CyberSecurity, Machine Learning, Deep Learning, Text Mining and Image Processing etc. Author has Presented and Published 3 Paper in Machine and Deep Learning