IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING APPROACH FOR IMPROVED NEURAL NETWORK SECURITY

[1]Rakshit Kothari, [2]Meenal Joshi, [3]Pankaj Kumar Vaishnav, [4]Mayank Patel, [5]Narendra Singh Rathore

[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor, [4]Head of Department, [5]Campus Director

[1]Department of Computer Science and Engineering,

[1]Geetanjlai Institute of Technical Studies, Udaipur, India

Abstract: Machine learning has moved from the lab to the forefront of operational systems during the past several years. Machine learning is used daily by Amazon, Google, and Facebook to enhance user experiences, recommend products, or help people connect socially through new applications and enable personal connections. The potent capability of machine learning is also present for cyber security. Machine learning will be used by cyber security to enhance malware detection, priorities events, identify breaches, and notify enterprises of security risks. Advanced threats and targeting, including organization profiling, infrastructure vulnerabilities, and potentially interconnected security breaches and exploits, can be detected using machine learning. The landscape of cyber security may be drastically altered by machine learning. As many as 3.75 million new samples can be represented by malware alone every hour. Conventional methods of malware analysis and detection are unable to keep up with evolving threats. Cyber-attacks are being delivered at frightening rates thanks to new attacks and clever malware that may avoid network and end-point detection. To combat the expanding malware problem, new methods like machine learning must be used. This study explains how cyber defense analysts may utilize machine learning to find and highlight sophisticated malware. The findings of our preliminary study are presented, and future research to advance machine learning is discussed.

Index Terms **- Data mining, dynamic analysis, malware detection, recurrent neural network, convolutional neural network.**

I. INTRODUCTION

In order to better understand the need for security employees globally, a Global Information Security Workforce Survey was carried out in 2019. Around 20,000 information security experts were polled as part of the study by the International Information Systems Security Certification Consortium (ISC) According to the workforce research, by 2023 there would be a shortage of more than 3.5 million security employees several businesses have discovered that there is a shortage of security personnel compared to the demand for cyber security services. In a recent interview, Symantec CEO Vincent Pilette predicted that by 2022, there will be a shortage of 3.75 million people worldwide. According to a Forbes prediction, the cyber security business will grow from $75 billion in 2015 to $370 billion in 2023. A Raytheon survey conducted in 2022 found that the demand for cyber security professionals is growing at a rate 3.5 times faster than the IT job market. Comparable investigations by Microsoft revealed that 55% of the examined firms were unable to fill critical cyber security positions. A majority of these businesses reported expecting a cyber-attack in the following 12 months in the same study. Most commercial, governmental, and academic institutions continue to face significant challenges from targeted assaults like malware and ransomware. Unless a" ransom" was paid, ransomware crippled local governments and healthcare institutions in 2020 by encrypting files. Ransom ware assaults [14], according to Symantec, increased by 65% in 2022. Developers of ransomware may soon rake in hundreds of millions of dollars thanks to these very lucrative attacks. From conventional IT targets to smart phones and other smart devices, ransomware migrated in 2022. In the very near future, Symantec is confident that ransomware will spread to IoT, wearable technology (such as smart watches), and smart homes.

The worldwide security community has an enormous problem in terms of being able to educate and deliver cyber security knowledge. These employees frequently require classroom, practical technological training, and operational knowledge to become competent. Nevertheless, the pipeline for education and training lacks the necessary volume to meet or even approach global demand in the coming years. Dealing with the millions of new threats that occur every day is a formidable job for organizations. These assaults come in a variety of sizes and employ a variety of threat vectors, such as ransomware, distributed denial-of-service attacks, polymorphic/metamorphic malware, common platform and network hacking, and email phishing. Several enterprise might be swiftly overwhelmed by these multi-vector attacks. The difficulty most security service and management firms face [8] [17]. Also, organizations are having problems with employee morale, psychology, and turnover of security professionals. The intelligence community as a whole and businesses are seeking for solutions to the constant bombardment of attacks on data, networks, systems, and mobile platforms. It is well acknowledged there is an enormous talent and ability gap in the world for cyber security. Knowing how the scarcity affects the commercial world, national security, law enforcement, and the intelligence

community as a whole. Figure 1 shows the structure of Neural Networks. Several experts have suggested using cyber security to combat organized crime, narcotics, terrorist acts, financial crime, corporate insider threats, espionage, and other dangers. When taking into account the Internet of Things (IoT) tsunami that is anticipated to impact the security environment by 2022, the outlook for fixing the holes appears to be less promising. The security industry as a whole is significantly impacted by these shortages. For IT companies, high school kids, and university students who want to work in cyber security, the acknowledged skills shortage offers a rare opportunity. Global cybercrimes including financial fraud, online child exploitation, and payment frauds are performed so often that they necessitate a round-the-clock response by international law enforcement organizations. An enormous number of cybercrimes, including those involving the Office of Personnel Management, Blue Cross/Blue Shield Anthem, Target, Home Depot, and Ashley Madison, necessitated the response of cyber security specialists in 2021–2022. By taking advantage of insufficient security, gaps in security design, and/or exploiting the vulnerabilities inside the IT infrastructure, adversaries have taken advantage of or hacked government and private computer systems [17].

This report makes the case that cutting back on the number of cyber security workers required to gather, evaluate, and disseminate information on malware detection gives a rare chance to reduce the cyber skills gap. A focus area for education and training in the realm of cyber security is provided by machine learning.

II. RELATED WORK

Using computer software, machine learning applications offer a novel method for addressing basic challenges in science and engineering. Deep learning has advanced significantly over the past 20 years and offers a simple entry point for individuals new to the topic. A rising number of commercial organizations are using machine learning in real applications since it transitioned from a laboratory ”black-box” setting. Computer vision, audio recognition, natural language processing, robot control, and other software applications have been created through machine learning. Machine learning is used by large corporations like Amazon, Google, and Facebook to enhance user experience, recommend purchases, and advertise special offers. Instead of writing the conventional input, machine learning engineers discover that creating samples of desired output makes training a system much simpler. Machine learning has had an impact on many businesses with data-intensive problems like cyber security. Similar prospects for development in domains like biology, cosmology, and social science have been provided by machine learning [4].

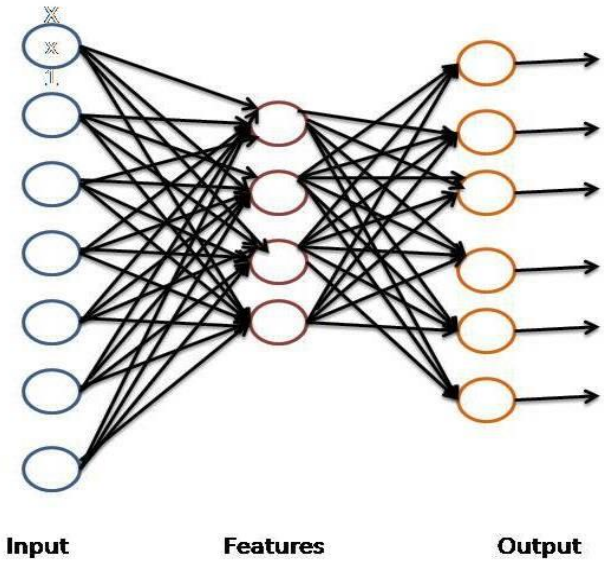

Fig. 1. Neural Network

Large amounts of experimental data may be processed and analyzed in unique ways using ml algorithm. Theoretically, machine learning algorithms can offer distinct perspectives on” big data” and can be improved to increase associated performance metrics in vertical applications. The algorithms used for machine learning can have a broad range of special features (e.g., decision trees, support vector machines, deep learning, neural networks and advanced clustering). The algorithms used for machine learning can have a broad range of special features (e.g., decision trees, support vector machines, deep learning, neural networks and advanced clustering). Yet, machine learning provides methods for analyzing large amounts of data in novel ways, allowing for the creation of evolutionary approaches and the improvement of optimization rough consecutive generations of algorithms [10].

This processing of massive volumes of data is best suited by machine learning and cyber security. Platforms and networks are frequently attacked.. With the variety of tools available for target scanning and evaluation, these assaults are more successful. Nowadays, adversaries use machine learning to improve their assaults. The majority of network security devices offer logs and other informational traces about unusual actions. Unfortunately, for the majority of security activities, these logs have low importance. To highlight occurrences for cyber security analysts, it is recommended that these logs be consumed by a Security Incident and Event Management (SIEM) system. However, sophisticated attackers conceal their presence and use deceptive tactics to avoid log management processing. In 2022, McAfee estimates that there will be 700 million new malware samples [2] [7]. Also, the sophistication of the virus makes it much harder for cyber security experts to identify increasingly advanced infections. It becomes increasingly difficult to notice incidents and alert cyber security analysts to them. To solve the rising and high volume issue, new alerting techniques must rely on technology rather than manpower. Technology will assist the labour force in

comprehending and responding to cyber threats. Machine learning provides some encouraging outcomes for handling security events after they happen.

III. AUTOMATED SECURITY EVENT REACTION

Modern security tools often provide system owners notifications and events based on monitoring with signatures and unusual activity. These occurrences typically reveal malicious activity and produce alerts at network devices like firewalls, intrusion prevention systems, and intrusion detection systems (IPS/IDS), as well as endpoint protection tools like anti-virus software, host intrusion protection systems (HIPS), host firewalls, and other tools. Correct operational and correlation of various events that are visible at the network and host levels are required for enterprise-wide cyber security cooperation. The majority of system owners are mostly worried about protecting systems and networks from known assaults. Yet, identifying zero-day attacks and sophisticated persistent threats is a challenge for both mature and advanced enterprises (APTs). Since they are ”low and sluggish” assaults that easily get buried in the ”noise” of millions of other events, these sorts of attacks are more challenging to detect. Security event management methods must be designed in an adaptable manner in order to respond to new and unique assaults while continuing to handle the known threats in order to meet the explosion of security events identified across businesses. Security event management methods must be designed in an adaptable manner in order to respond to new and unique assaults while continuing to handle the known threats in order to meet the explosion of security events identified across businesses. Situational awareness and gradual learning are required for new event management tactics. As a response to identified events, today’s Security Management Centres (SMC) and Distributed System Centres (DSC) create an attack context. Information security analysts compile alerts/incidents from sources such packet capture, net flow data, and device logs that also include attack forensic event information. Implementing security event management processes requires the assembly of ontologies and logic that can be shared throughout a security infrastructure, as well as contextually and semantically rich detection information [5] [9] [12]. To encourage adaptive learning for security incidents, new solutions must connect these systems and share information. In researching security event management there are a number of approaches that enable adaptive learning.

A technique to analysing categorical similarity measurements based on a numerical taxonomy is provided by Sneath and Sokal. Sneath and Sokal give a strategy for problem-solving and help create solutions based on category similarity and relevance. Their study showed that by comparing the similarity of category elements, traits, or qualities, issues may be categorized by numerical and categorical analysis [1] [11]. Moreover, qualities may be assessed to establish the importance of a feature. Nevertheless, because they are often binary in nature, the issues this technique resolves do not lend themselves to complicated issues like security incidents. Given that the majority of security incidents are binary either good or bad the information needed to determine them is highly varied and heterogeneous in character. Wilson and Martinez provide a method for the analysis of heterogeneous events using distance functions that classify occurrences according to continuous and categorical features. This method offers classes of information supervised learning, where each class contains a set of categorical and continuous qualities that may be assessed by distance metric learning. This strategy is quite helpful for datasets that are tiny and sparse. So far, there are a lot of data volumes in security event management that need to be examined. For numerous and substantial datasets, several learning algorithms have been created. Recurrent Neural Networks (RNN) and Convolution Neural Networks (CNN) provide a powerful approach for analysing diverse and large datasets. RNN and CNN also offer significant benefits for complicated and abstract data input layers. RNN and CNN are frequently used successfully for image processing, audio recognition, and natural language processing. Until now, it might be difficult to use RNN and CNN to address real-world problems like security event management. RNN and CNN offer a broad method for resolving frequent issues. Many problems and practical considerations surround the training of RNN and CNN models [3]. It was necessary to specify and label parameters for a sizable portion of the anticipated dataset when training RNN and CNN.

The training of RNN and CNN models is fraught with issues and practical concerns. For training RNN and CNN, it was required to specify and label parameters for a significant fraction of the predicted dataset. The many scenarios that the model anticipates must be covered during the training phase. Covering the majority of the projected data options for complex datasets can be quite difficult and time-consuming. The training process must attempt to utilize convex optimization techniques to solve no convex data model problems. RNN and CNN solve non-convex optimization problems by breaking the bigger problem down into smaller components. This approach is based on the idea that local convex optimization is more suitable for smaller parts. Convex optimization techniques are used to address the no convex issues, which increases performance overall when local improved components are put together. Often times, optimization cannot be fully performed and other approaches must be utilized to optimize the DNN model. To deal with outlier problems, ad hoc techniques like gradient trimming or batch normalizing are required [13]. Ad-hoc procedures are sometimes required to produce a cohesive model. Ad hoc techniques must, however, adjust the network RNN or CNN) utilizing back- propagation as a training component. It is difficult to adjust the RNN or CNN for complicated problems. Lack of data during RNN or CNN model training might lead to overfitting problems. For this, the practitioner must comprehend and build the RNN and CNN model with the appropriate levels of complexity. Thus, the practitioner must be an authority on the subject at hand and have adequate data coverage for the issue at hand. In our study, we describe the methodology and outcomes of creating CNN for large-scale security incidents.

IV. RESEARCH APPROACH

This study’s objective was to ascertain how machine learning may be used to address the security event management problem. Millions of warnings are sent to many corporate clients each day. According to McAfee, up to 4.5 million new malicious files are released every hour. Cyber defence analysts and security operations are quickly overwhelmed by the number of malware and network detections [16]. Due to the above mentioned scarcity of security experts, this issue cannot be resolved. To automate and coordinate the processing of security incidents, technology must be used. Machine learning in particular can give a way to deal with the quantity and inherent knowledge offered by cyber defence analysts. A fast analysis of real figures will give you a comprehensive understanding of the difficulties that SMC and DSC professionals deal with on a daily basis and will help to comprehend the challenges facing today’s cyber defence experts. A medium to big SMC/DSC, as seen in Fig. 1. The figures

presented here are normal for a mid-size company with less than 100,000 subscribers. The occurrences might potentially take over 5,500 man-days of work to clear or assess, which further emphasizes the necessity for automation. 16 analysts work for this firm for the SMC/DSC. This group has a daily capacity of 86 to 154 hours. There are more than 2,500 hours missing. There is a clear demand for automation, thus in collaboration with our customers, we have devised a method for explaining how to apply machine learning in a real-world setting. There are six steps in our research methodology which are as follows:

A. Increase business knowledge

The team convened with key players to go over the challenges and problems. The conference produced the following findings and issues that the machine learning project has to solve:

- The volume of daily notifications is overwhelming security analysts.
- Some warnings are routine and lack context or importance.
- Leadership is aware of the nature of the alerts, a method for reviewing the "unknown" or "unconnected" warnings must exist.
- There is no simple way to priorities and react to alarms.

B. Data and data dependency analysis

To gain a better understanding of the data, procedure, and results, the team gathered representative alerts and examined prior work. That turned out to be a considerably greater task than expected. The over 15,000 notifications that make up the universe of events. Formerly, these warnings were divided into nearly 18 categories. Moreover, certain events have low densities, making it unnecessary to replicate unique events. It quickly became apparent that there were over 15,000 notifications, much too numerous. These warnings were initially divided into 18 categories.

C. Work with subject-matter specialists

The category and sub-categories were reviewed and validated by subject-matter experts (SMEs). The actions done, independent and dependent alerts, and the priorities related to alert types and subcategories are also discussed by SME. Businesses might assess the quantity and frequency of warnings. It was also used to give the different notifications priority and weighting. After carefully weighing more than 10,900 warnings, the weights were analysed to determine their minimum and maximum values as well as their average, median, and min-max value. Also, it is important to be extremely explicit about the project's intended results.

D. Build a dataset

Data preparation may need between 85 to 95 percent of the time spent on data engineering, according to research. It is crucial to remember that in order to provide high-quality results, the dataset utilized for machine learning has to be examined, validated, and occasionally changed [6]. The dataset gives the different RNN and CNN algorithms the necessary input to build the learning model. Understanding the distribution of data and the range of values included in the dataset, as well as searching for missing or incomplete data, are all necessary steps in this process. In figure 2 represents the dataset generator of Tensor Flow libraries.

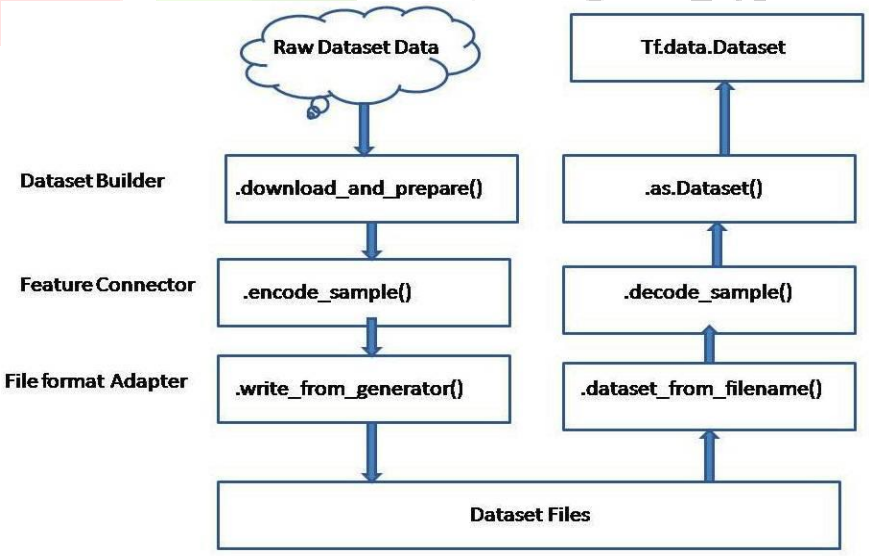

Fig. 2. Dataset Generator

E. Build a model

Using the supplied dataset, the team created the model. In order to include the dataset into a single model, many runs were performed. The model was developed and tested using Tensor Flow. Because of its distributed design, Tensor Flow was chosen for this study. To swiftly train and create sophisticated RNN and CNN models, Tensor Flow can effectively make use of hundreds of servers. Via parallelization, Tensor Flow has a special capacity to supply and control both computation and state management. As a result, the team was able to run several tests and experiments on a single dataset [15]. Figure 3 shows the architecture of Tensor Flow technique.

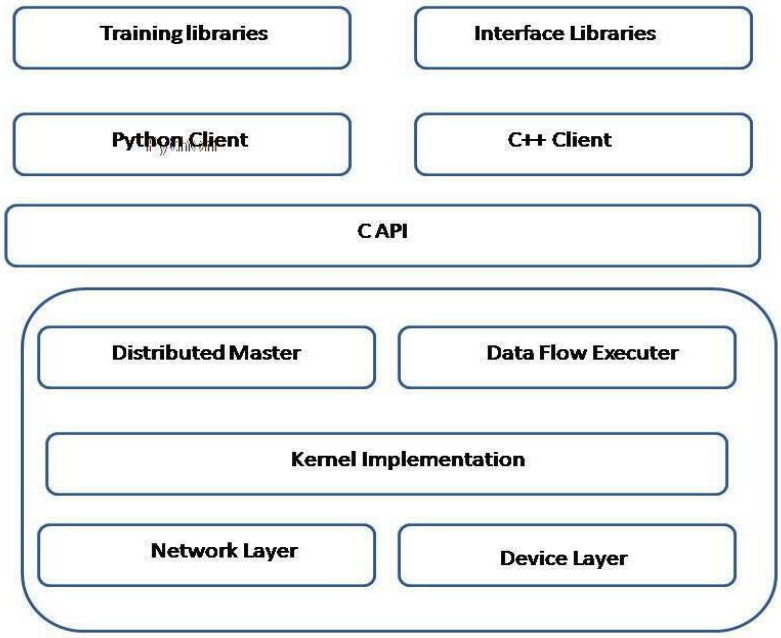

Fig. 3. Architecture of Tensor Flow

F. Analyse the model

Based on the performance of the classifier, the model developed using the test dataset is tested or evaluated. On the basis of standard stratified k-fold cross-validation, classifier performance is evaluated. Stratified cross-validation investigates the distribution of classes and whether they are distributed uniformly across each wave. Evaluation of the classifier's performance in accurately identifying the data instances from the test dataset is the aim of cross-validation. Evaluation of the dataset's correctly labelled instances is part of the model validation process.

V. CONCLUSION

This study's objective was to get a deeper understanding of how machine learning may be used to categorize various security incidents and warnings. The experiment went through each step of creating goals and objectives, gathering data, verifying the dataset, and creating a neural network model in a methodical manner. The experiment's ultimate objective was to determine if the model would appropriately respond to security incidents by notifying SMEs, notifying analysts, or notifying analysts while creating reports, depending on the severity of the security event. The accuracy of the model in carrying out these tasks was extremely good (95%). The neural network model for this study was created using Tensor Flow. The model accurately detected and responded to the approximately 13 million security events given in the test dataset. According to an analysis of the experiment's findings, implementing machine learning into SMC/DSC processes might cut down on security analyst time by 86%. With the use of machine learning, the initial issue of more than 3500 hours per day may be drastically decreased to 655. To further minimize the time required for responding to major security incidents, future efforts should aim to provide more help for cyber defence analysts.

REFERENCES

[1] Dai, H. L. (2015). Imbalanced protein data classification using ensemble FTM-SVM. IEEE transactions on nanobioscience, 14(4), 350-359, 2015.
[2] Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. Sensors, 21(11), 3758.
[3] Menzies, T., Greenwald, J., & Frank, A. (2006). Data mining static code attributes to learn defect predictors. IEEE transactions on software engineering, 33(1), 2-13.
[4] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 250, 113-14.
[5] Giri, K. C., Patel, M., Sinhal, A., & Gautam, D. (2019, April). A novel paradigm of melanoma diagnosis using machine learning and information theory. In 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE) (pp. 1-7). IEEE.

[6] Maurya, V. K., Mehra, R. M., & Mehra, A. (2016). Design and Analysis of Energy Efficient OPAMP for Rectifier in MicroScale Energy Harvesting (Solar Energy). In Proceedings of the International Congress on Information and Communication Technology: ICICT 2015, Volume 2 (pp. 229-240). Springer Singapore.
[7] Huang, G., Song, S., Gupta, J. N., & Wu, C. (2014). Semi-supervised and unsupervised extreme learning machines. IEEE transactions on cybernetics, 44(12), 2405-2417.
[8] Kothari, R., Choudhary, N., & Jain, K. (2021). CP-ABE Scheme with Decryption Keys of Constant Size Using ECC with Expressive Threshold Access Structure. In Emerging Trends in Data Driven Computing and Communications: Proceedings of DDCIoT 2021 (pp. 15-36). Springer Singapore.
[9] Li, H., Chung, F. L., & Wang, S. (2015). A SVM based classification method for homogeneous data. Applied Soft Computing, 36, 228-235.
[10] Parvin, H., Minaei-Bidgoli, B., & Alinejad-Rokny, H. (2013). A new imbalanced learning and dictions tree method for breast cancer diagnosis. Journal of Bionanoscience, 7(6), 673-678.
[11] G. Kaur and H. Singh, "Data Mining Techniques for Text Mining", Indian Journal of Science and Technology, vol. 9, no. 44, 2016.
[12] Rathore, R., Sharma, R., Bhanawat, R., Soni, P., Soni, P. R., & Sachdev, P. (2022). Inter-linked platform for Campus Placement in Higher Educational Institutions of India. International Journal of Advanced Research in Computer Science, 13.
[13] Khan, F., Kothari, R., & Patel, M. (2022). Advancements in Blockchain Technology With the Use of Quantum Blockchain and Non-Fungible Tokens. In Advancements in Quantum Blockchain With Real-Time Applications (pp. 199-225). IGI Global.
[14] Khan, F., Kothari, R., Patel, M., & Banoth, N. (2022, April). Enhancing non-fungible tokens for the evolution of blockchain technology. In 2022 International conference on sustainable computing and data communication systems (Icscds) (pp. 1148-1153). IEEE.
[15] Vaishnav, P. K., Sharma, S., & Sharma, P. (2021). Analytical review analysis for screening COVID-19 disease. International Journal of Modern Research, 1(1), 22-29.
[16] Kumari, R., & Vaishnav, P. K. (2022). Hybrid Implementation System for Heart Disease Prediction in Data Mining". In 2022 Journal for Basic Sciences.DOI:10.37896/JBSV22.9/1454, Page No:37-46.
[17] Vyas, H., Vaishnav, P. K., (2022). Implementation of Machine Learning and Artificial Intelligence In Bioinformatics. Using Cnn.DOI:20.18001.GSJ.2022.V9I9.22.40124.