International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 112-118
Published Online February 2021 in IJEAST (http://www.ijeast.com)

PREDICTION OF NETWORK ATTACKS USING MACHINE LEARNING TECHNIQUES

Dr. G. Umarani Srikanth
Professor, CSE department,
Bharath Institute of Higher Education and Research, Chennai

Priyadharsini.S
St Peter's College of Engineering and Technology, Chennai

Abstract: **The networked systems become more and more pervasive and businesses still acquire a lot of sensitive data online, so that the quantity and class of cyber-attacks and network security breaches has risen dramatically. There are also instances that so many volumes of data are hacked even without the knowledge of the people concerned. So far setting an Intrusion Detection System (IDS), it is obvious to set the true working environment to model the possibilities of attacks. Therefore, it is imperative to design a software that will be able to identify network intrusions, in order to protect a computer network from the unknown users. For overcoming this challenge, it is essential to predict whether the connection is targeted or not from KDDCup99 dataset utilizing machine learning techniques. The objective of this work is to investigate machine learning based algorithms for enhancing packet connection transfers forecasting using ensemble learning voting classifier techniques. It is proposed to deploy AI-based technique to precisely anticipate the DOS, R2L, U2R, Probe and large assaults. Results showed that the viability of the proposed AI calculation strategy can be contrasted and the best exactness with accuracy, Recall and F1 Score.**

Keywords: Artificial intelligence (AI), Machine Learning (ML), classification method, python, Prediction of Accuracy result.

I. INTRODUCTION

The networking systems automatically track their networks traffic and report suspecting or doubtful behavior, work in one of either styles: misuse detection and anomaly detection. Misuse detection method go in search of precise signatures of a known malicious behavior while anomaly detection will try to form a model for what it contains normal network traffic patterns , then flag deviations from those patterns. Due to this, signature-based antivirus software is becoming a non-entity and it is considered as inconvenient with spoofing of signatures and ignorant latest sophisticated attacks. Because of this, it struggles to remain as a dominant force in latest threat arena. The anomaly-based intrusion detection gives the idea of getting the power to find novel attacks even before they have been understood and characterized by security analysts also as having the power to identify the variations on existing attack methods.

For creating data in the Intrusion Detection System (IDS), there is need to line the important working condition to find all the possible type of attacks, it is imperative to examine, transform and model the data. The data quality zeroes on the uprightness and reliability of knowledge obtained and used in an evaluation. Data quantity engages with the amount of knowledge obtained for the validation. The job needs different ground truth databases in its region and thus the practicability could be finished efficiently if the data quality and features of for the precise are good. Image processing, web site analysis, and other things have the standard and permitted ground truth databases for evaluation. Similarly, lot of the PC network intrusion detection systems use the KDD Cup99 to classify and analyze network traffic and it clearly depicts the formation of KDDCup99 dataset and its features.

Eventually Machine learning has evolved to predict the long run data from the past data. The activity of coaching and prediction involves usage of very special algorithms. Machine Learning feeds the training data to an algorithm, and thus the algorithm utilizes this training data to supply predictions on a replacement test data. Machine learning algorithms are often divided into three types. These are supervised learning, unsupervised learning and reinforcement learning. It also provides the training algorithm, which evaluates the clustering of the input data. At last, the reinforcement learning dynamically interacts with its environment and it also receives good or negative feedback to reinforce its performance. In machine learning and statistics, classification is based on supervised learning approach during which the program learns from the information input given. Then it utilizes this learning to classify new observation.

II. LITERATURE SURVEY

In this paper [1] the authors discussed about the Cyber- Physical Systems (CPSs), which is the cutting-edge period of planned system during which preparing, correspondence, and control progressions are significantly joined. Assessment on CPSs done

on a fundamental level were noteworthy for structured systems in various huge application spaces like transportation, imperativeness, and clinical structures. The authors surveyed CPS research from both with a recorded viewpoint with reference to progressions made from early times control structures. Results on CPSs are recorded in various material assessment spaces, as an example, organized control, hybrid structures, persistent enrolling, steady frameworks organization, remote sensor frameworks, security, and model-driven new development.

In this paper [2] the author has discussed that by increasing the instances of privacy leakage in CPSs and the corresponding intense consequences, have caused unimaginable worries in overall population. In most security protecting instruments, it is proposed to ensure tricks to the individual information, so that structure execution is subverted at an equivalent time. They took in to consideration the trade-off between particular security and structure execution within the CPSs. The authors also figured the introduction improvement issue subject to a given differential assurance that is needed. Reenactment results were given to visualize the proposed part, which modifies the trade-off between system execution and assurance. They also recognized future investigation subjects on the insurance sparing issue in CPS.

In this paper [3] the authors highlighted that by changing the communication base between sensors and control systems, aggressors achieved the safety adroit system structures with the usage of strategies like DoS attack and other types of possible attacks. They presented a numerical model of the structure in order to propose a tremendous security framework. They presented an innovative mathematical model of the system to learn the pitfalls and suggested a suitable security model for the designing the smart grid. The $\chi 2$-discoverer is an exhibited convincing exploratory procedure used with the Kalman channel for the estimation of the association between subordinate components and a movement of pointer factors. The $\chi 2$-locator can recognize structure insufficiencies like DoS attack and long stretch self-assertive attacks. The authors have validated the limits of the $\chi 2$-detector in finding the statistically obtained False Data Injection attack and suggested new Euclidean detector to demonstrate the efficacy of the proposed approaches using extensive simulations and testing on practical systems.

Here authors [4] have discussed that advanced physical structures are inevitable in power systems, transportation frameworks, present day control structures, and essential establishments. These structures got to work constantly despite unforeseen dissatisfactions and external pernicious attacks. They proposed a logical structure for computerized physical systems, attacks and screens. They also depicted vital watching requirements from system speculative and description theoretical perspectives. They made arrangement bound together with scattered ambush disclosure and conspicuous evidence screens.

The main aim of this paper [5] is to form the model-based techniques that are able to detect integrity attacks. The impact of respectability assaults on the control frameworks is to break down and to equip for uncovering such proposed assaults. The objective of this work lies in identifying the states of the attainability of the replay assault and recommending to improve the likelihood of identification by surrendering control execution.

In this paper [6] the authors marked a turn round within the advent of the primary cyber warfare armory , called as Stuxnet. It was not about industrial espionage: it did not steal, manipulate, or erase information.

The authors [7] expressed that the industrial system communication networks are susceptible to reconnaissance, response injection, command injection, and denial of service attacks which will lead misfortune circumstances for control framework administrators who deploy these administrations. This paper portrayed tons of 28 digital assaults against mechanical control frameworks that utilize the MODBUS application layer which organizes the convention. The paper additionally expressed tons of independent and state-based interruption recognition framework rules which may be utilized to identify digital assaults and to store proof of assaults for post-occurrence investigation.

In this paper [8] the authors discussed that in mechanical plants, as an example, atomic force plants, framework tasks are performed by implanted controllers arranged by supervisory control and data acquisition programming. Such malware assaults can cause huge expenses to the association for recuperation, cleanup, and maintenance movement. SCADA frameworks in operational mode produce colossal log records. These records are valuable within the investigation of plant operation and diagnostics during a progressing assault. This paper investigated techniques and calculations to create a successful observing plan against control mindful digital assaults. It additionally clarified complicated calculation procedures, for example, the computational geometric technique and least-squares estimation which may achieve success in screen plan.

In this paper [9] the authors analyzed the issues of prognostics and health, the board for atomic force frameworks, innovations in other mechanical application zones. The algorithms to model that evaluate the degradation state of reactor components. The algorithms that validated the information of degradation in order to find Remaining Useful Life (RUL) and Probability of Failure (POF) of the rector component were also studied. The prognostic results were used for the management of evolving health and condition of nuclear plant Systems, Structures, and Components (SSCs). It was found that the POF information was

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 112-118
Published Online February 2021 in IJEAST (http://www.ijeast.com)

utilized in a Probabilistic Risk Assessment (PRA) model which is helpful to evaluate the risk exposed of the degradation and the corresponding reduced safety margin. Estimated values of RUL and its uncertainties can also be utilized to give valuable input to plant engineers for Operations and Maintenance (O&M) planning or in automated optimal control algorithms.

Multiple sclerosis (MS) is a chronic disease, this illness affects the insulating cover of nerve cells in central neural system. In this journal[10] the authors presented the work to identify Multiple Sclerosis issues from sound controls in attractive reverberation imaging. The Multiple sclerosis imaging information was obtained from the eHealth research facility, and thus Healthy Controls (HC) imaging information was examined. Between examinations, standardization was utilized to evacuate the dim-level contrast. The authors modified the misclassification expenses to lighten the impact of uneven class appropriation to the grouping execution. Two-level fixed wavelet entropy was used to obviate highlights from mind pictures. Results were verified using three AI-based classifiers: the selection tree and K-Nearest Neighbors(KNN). The exploratory outcomes indicated the KNN played out among the two classifiers.

In this paper [11] the authors discussed that while both cost-touchy learning and web-based learning are concentrated independently, these two issues have only from time to time been attended at an equivalent time. This paper researched a category of algorithmic methodologies reasonable for online cost-delicate learning, intended for solving such issues. It was proposed to use existing techniques for online group calculations and consolidate these with clump mode strategies for cost-touchy sacking/boosting calculations. Inside this structure, the authors portrayed a couple of hypothetically solid online cost-touchy packing and online cost-delicate boosting calculations and showed that the assembly of the proposed calculations ensured under specific conditions.

III. OVERVIEW OF THE SYSTEM

Because of the enormous volumes of information which are so complicated, the dynamic properties of interruption practices, and information mining-based Intrusion Detection procedures have been utilized to arrange traffic information. The ongoing enhancement in PC innovation, a lot of information could be gathered and put away.

AI methods can also be helpful in the mix of PC based models in the system condition giving chances to encourage and improve crafted by organize security specialists. It automatically improves the efficiency and originality of information and data. System Intrusion Detection targets recognizing the conduct of the system. This work represents the execution of four administered learning calculations, C4.5 Decision tree Classifier (J48), Instance-Based Learning (IBK), Naive Bayes (NB), and Multilayer Perceptron (MLP) in WEKA condition, in an Offline situation. The classification models were trained utilizing the information gathered from Knowledge Discovery Databases for Intrusion Detection.

The frameworks that are formed are all utilized for anticipating the danger of the assaults in a web server condition. The prediction accuracy of the classifiers was assessed utilizing K-crease cross-validation and outcomes have been contrasted with acquiring the precision. It needs to discover accuracy of the preparation dataset, accuracy of the testing dataset, specification, false positive rate, exactness, and review by looking at calculation utilizing python code.

The steps involved in Building the data model is depicted in the Fig.1.

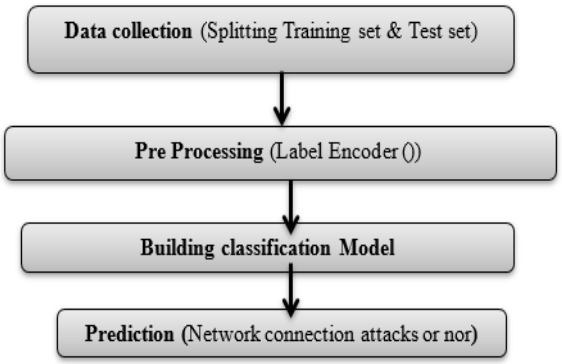


Fig:1 data flow diagram for Machine learning model

IV. PROPOSED SYSTEM

Machine learning and statistics, use a supervised learning approach for classification. Here the computer program studies from the data input fed in to it and then utilizes these studies to classify newly observed things.

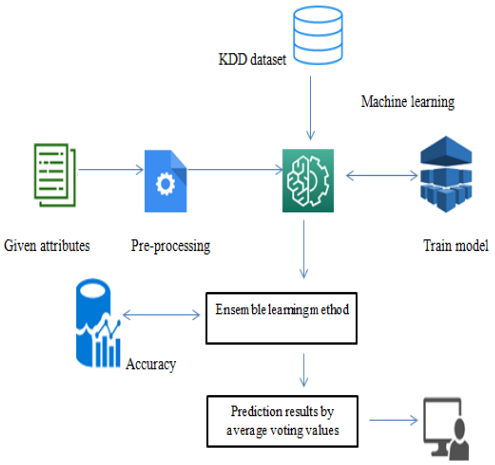


Fig.2 architecture diagram

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 112-118
Published Online February 2021 in IJEAST (http://www.ijeast.com)

This set of data's will be bi-class or a multi-class. In supervised studies , algorithms understand from labeled data. After understanding, the algorithm identifies the suitable data to be additional to new data which is depends on pattern and then linking the patterns to the unlabeled new data as shown in Fig.2.

V. LOGISTIC REGRESSION

This is a strategy that can be measured for obtaining an informational collection in which at least one autonomous factor decides result and the result is evaluated with a dichotomous variable. The main aim of strategic relapse is to identify the suitable modeling to illustrate the relation between the dichotomous attribute of intrigue (subordinate variable = reaction or result variable) and a lot of autonomous (indicator or informative) factors. Calculated relapse is a machine learning characterization calculation which is used to find futuristic likelihood of a straight outward variable. In strategic relapse, the needy variable is a parallel variable that contains information coded as 1 (truly, achievement, and so forth.) or 0 (no, disappointment, and so on.).

1. Logistic regression Assumptions:
2. For a binary logistic regression the dependent variable should be binary.
3. For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
4. Meaningful variables must only be included.
5. The independent variables must be independent of each other.
6. Independent variables must be linearly related to the log odds.
7. Logistic regression needs large sample sizes.

VI. K-NEAREST NEIGHBOR (KNN)

K-Nearest Neighbor is a managed AI calculation that stores all occurrences related to preparing information which zeroes in on n-dimensional space. At the place when obscure discrete information is received , it evaluates the nearest k number of examples spared (closest neighbors) and returns the often recognized class as the forecast and for truly esteemed information, it gives back the mean of k-closest neighbors. When calculating weighted closest neighbor, it loads the commitment of every one of the k-neighbors as per their separation utilizing the accompanying question giving more noteworthy load to the nearest neighbors.

For the most part, KNN is powerful tool since it is averaging the k-closest neighbors. It makes expectations about the approval set utilizing the whole preparing set. KNN makes a forecast about another example via looking through the whole set to discover the k "nearest" occasions. "Closeness" is resolved utilizing a vicinity estimation overall highlights.

VII. RANDOM FOREST

Irregular timberlands or arbitrary choice backwoods are a troupe learning technique for arrangement, relapse, and various undertakings, that work by building a huge number of choice trees at preparing time and yielding the class that is the method of the classes (characterization)[12]. Arbitrary timberland is a kind of managed AI calculation dependent on outfit learning. Troupe learning is a kind of realizing where you join various sorts of calculations or a similar calculation on different occasions to frame an all the more impressive forecast model. The irregular woods calculation joins different calculations of a similar sort for example different choice trees, bringing about backwoods of trees, consequently the name "Arbitrary Forest". The arbitrary woodland calculation can be utilized for both relapse and grouping undertakings.

These are the essential ways used in performance of the random forest algorithm:

- Pick N random records from the dataset.
- Construct a choice tree supported these N records.
- Select the number of trees you would like in your algorithm and repeat steps 1 and 2.
- If there is an issue of a regression problem, for a replacement record, each tree within the forest finds a worth for Y (output). the ultimate data is often calculated by taking the standard of all the data's predicted by all the trees in forest. Or, just in case of a classification problem, each tree within the forest predicts the category to which the new record belongs and majority vote goes to the new record.

VIII. TESTING AND IMPLEMENTATION

- Download and install anaconda and get the most beneficial package for machine learning in Python.
- Load a dataset and learn its structure using statistical summaries and data visualization.
- In the Machine learning models, choose the best and take in to confidence that the accuracy can be trusted.

Python is a mainstream and amazing deciphered language. In contrast to R, Python is a finished language and stage that you can use for both innovative work and creating creation frameworks. There are likewise a great deal of modules and carry out every responsibility.

IX. PERFORMANCE MEASUREMENTS OF ML ALGORITHM

The results of predictions of attacks using various machine learning algorithms are shown in tables 1,2,3,4 and 5 and the corresponding comparison graphs are drawn shown in Fig.3 to 7 respectively.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 112-118
Published Online February 2021 in IJEAST (http://www.ijeast.com)

Table No:1 DOS Attack Prediction

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.5	1	1	1	1	1
Recall	1	1	1	1	1	1
F1-Score	1	1	1	1	1	1
Sensitivity	1	1	1	1	1	1
Specificity	0	0	1	0	0	1
Accuracy (%)	92.44	91.5	99.99	97.2	94.2	90.06

It was observed that the highest accuracy for DoS attack is Random Forest algorithm.

Table No:2 R2L Attack Prediction

Parameters	**LR**	**DT**	**RF**	**SVC**	**KNN**	**NB**
Precision	1	1	1	0.97	1	0.96
Recall	1	1	1	1	1	1
F1-Score	1	1	1	0.99	1	0.98
Sensitivity	0.99	0.97	0.98	0.95	0.93	0.91
Specificity	0.99	0.99	0.99	0.99	0.99	0.99
Accuracy (%)	99.98	93.6	94.44	93.56	98.97	95.26

It was found that the highest accuracy for R2L attack is Logistic Regression

Table No:3 U2RAttack Prediction

Parameters	**LR**	**DT**	**RF**	**SVC**	**KNN**	**NB**
Precision	0.99	0.99	0.99	0.99	0.99	1
Recall	0.99	0.99	0.99	0.99	0.99	0.98
F1-Score	0.99	0.99	0.99	0.99	0.99	0.99
Sensitivity	0.99	0.95	0.92	0.90	0.93	0.99
Specificity	0.69	0.70	0.67	0.67	0.65	0.99
Accuracy (%)	98.68	97.61	98.51	98.64	98.61	98.67

It was found that the highest accuracy for U2R attack is Logistic Regression and Naive Bayes algorithms.

Table No:4 Probe Attack Prediction

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	1	1	1	1	1	1
Recall	1	1	1	1	1	0.99
F1-Score	1	1	1	1	1	1
Sensitivity	0.98	0.99	0.99	0.99	0.99	0.99
Specificity	0.98	0.98	0.97	0.90	0.98	1
Accuracy (%)	99.89	99.88	99.90	99.82	99.90	99.11

It was observed that the highest accuracy for Probe attack is Random Forest and KNN algorithm.

Table No:5 Overall Network Attack Prediction:

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.85	0.93	0.93	0.89	0.93	0.84
Recall	1	0.95	0.95	0.94	0.94	1
F1-Score	0.92	0.94	0.94	0.92	0.93	0.91
Sensitivity	0.99	0.94	0.94	0.94	0.94	0.99
Specificity	0.97	0.98	0.98	0.98	0.98	0.97
Accuracy (%)	97.78	98.4	98.44	97.80	98.32	97.61

It was found that the highest accuracy for overall Network attack is Random Forest algorithm.

X. PERFORMANCE MEASUREMENTS OF ML ALGORITHMS

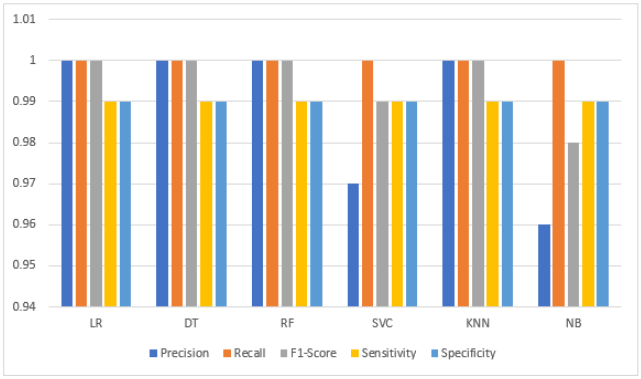


Fig.3 Dos Attack Prediction

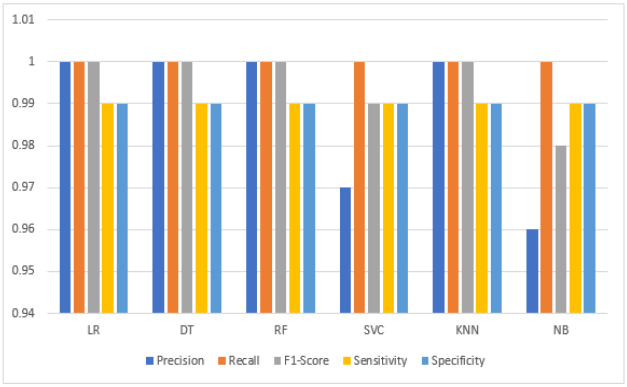


Fig.4 R2L attack prediction

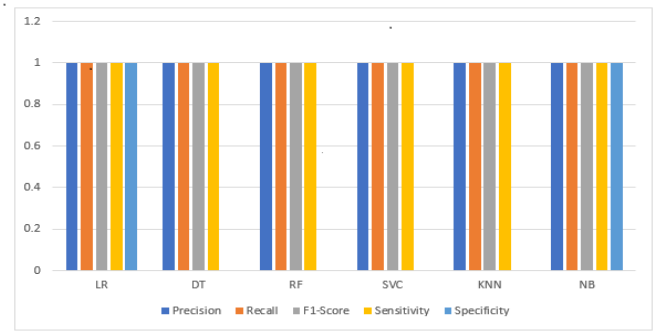


Fig.5 U2R attack prediction

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 10, ISSN No. 2455-2143, Pages 112-118
Published Online February 2021 in IJEAST (http://www.ijeast.com)

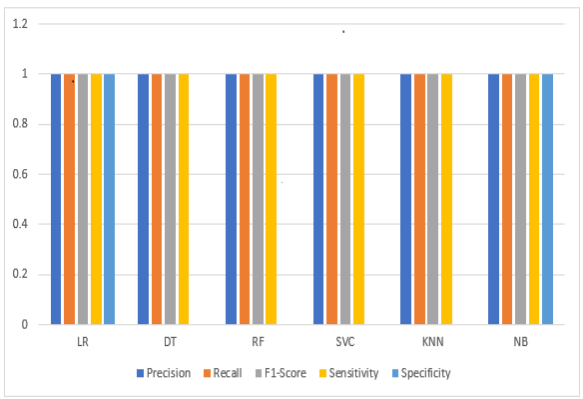


Fig.6 probe attack prediction

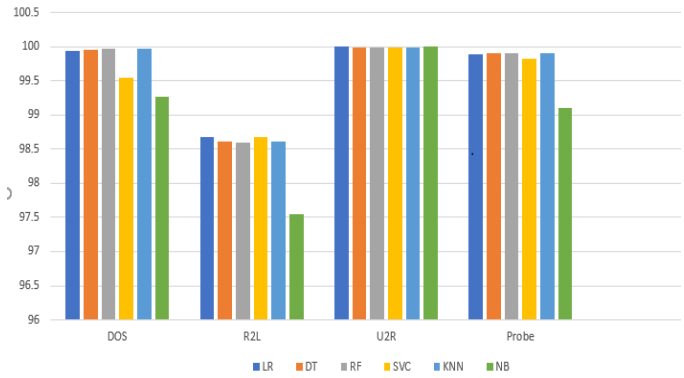


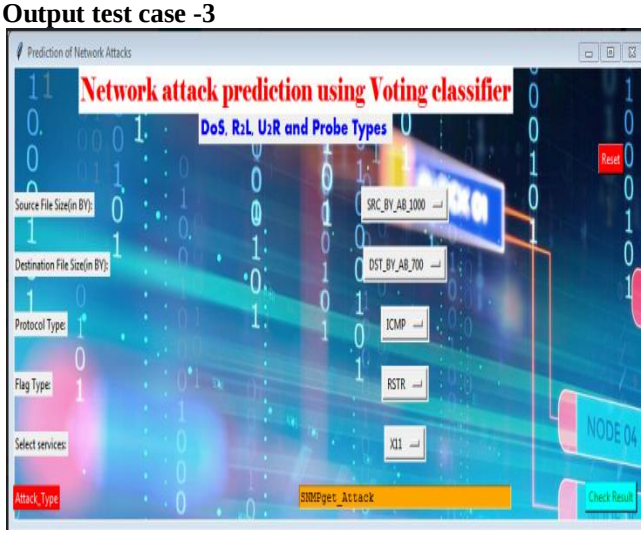
Fig.7 overall network attack prediction

Snapshot of the input

Output :Test-01

➢ **output – test 02**

Output test case -3



XI. CONCLUSION

This paper concludes that in order to present a prediction model with the help of machine learning techniques for strengthening over human accuracy and supply with the scope of early detection. It can also be inferred from this model that usage of machine learning technique is beneficial in developing prediction models which may help to network sectors in reducing the process of diagnosis to eradicate human errors. The work executed and implemented, deployed AI-based techniques to precisely anticipate the networks attacks and large assaults. Results have shown that the usage of the proposed AI calculation strategy can be used to achieve fair accuracy, Recall and F1 Score.

XII. REFERENCES

[1] Kim K D., and Kumar P. (2012). Cyber–physical systems: A perspective at the centennial, in Proc.IEEE, vol.100, no.13, pp.1287–1308.
[2] Zhang H., Shu Y., Cheng P., and Chen J.(2016). Privacy and performance trade-off in cyber-physical Systems, IEEE Network, vol. 30, no. 2, pp. 62–66.
[3] Manandhar K., Cao X., Hu F., and Liu Y.(2014). Detection of faults and attacks including false data injection attack in smart grid using Kalman filter, IEEE Transactions on Control of Network Systems, vol.1, no.4, pp.370–379.
[4]Pasqualetti F,. D¨orfler F., and Bullo F.,(2013). Attack Detection and Identification in Cyber-Physical Systems, IEEE Transactions on Automatic Control, vol.58, no.11, pp.2715–2729.
[5]Jia QS., Shi L., Mo Y., and Sinopoli B., (2012). On optimal partial broadcasting of wireless sensor networks for kalman filtering, IEEE Transactions on Automatic Control, vol.57, no.3, pp.715–721.
[6]Langner R.,(2011). Stuxnet: Dissecting a cyber warfare weapon, IEEE Security & Privacy, vol.9, no.3, pp.49–51.
[7]Gao W., and Morris TH., (2014). On cyber attacks and signature based intrusion detection for modbus based industrial control systems, Journal of Digital Forensics, Security and Law, vol. 9, no.1, pp.37-55.
[8]Gawand HL., Bhattacharjee A., and Roy K.,(2017), Securing a cyber physical system in nuclear power plants using least square approximation and computational geometric approach, Nuclear Engineering and Technology, vol.49, no.3, pp. 484–494.
[9]Coble J., Ramuhalli P., Bond L., Hines J., and Upadhyaya B.,(2015). A review of prognostics and health management applications in nuclear power plants, International Journal of Prognostics and Health Management, 6 (Special Issue Nuclear Energy PHM) 016, 22.
[10]Zhang Y.,Lu S., Zhou X., Yang M.,Wu L., Liu B., Phillips P., and Wang S., (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. Simulation, vol.92, no.9, pp. 861–871.
[11]Wang B., and Pineau J., (2016), Online bagging and boosting for imbalanced data streams. IEEE Transactions on Knowledge & Data Engineering, vol.28, pp. 3353-3366, vol. 28.
[12]Jaiswal JK., and Samikannu R.,(2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression, in Proc. WCCCT,pp.65-68.