Ransomware Detection and Classification Using Predictive Analytics

Jalaja.S, Prabakaran.M, Vishal.T, Ravi Teja N.G

Department of Electronics andCommunication Engineering
Vel Tech High Tech Dr.RangarajanDr.Sakunthala Engineering, College
Chennai, Tamil Nadu, India.

Department of Electronics and Communication Engineering
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering, College
Chennai, Tamil Nadu, India.

Department of Electronics and Communication Engineering
Vel Tech High Tech Dr.RangarajanDr.Sakunthala Engineering, College
Chennai, Tamil Nadu, India.

Department of Electronics and Communication Engineering
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering, College Chennai, Tamil Nadu, India.

,
ngravi13@gmail.com, Jalaja@velhightech.com, prabakarann649@gmail.com, vishalkeerthi4568@gmail.com

Abstract— The capacity for ransomware groups to substantially impair computer systems, data centres, websites, and mobile apps in a number of businesses and professions makes them serious security risks for cybersecurity. Conventional anti-ransomware software developers struggle to counter newly created, complex threats. Therefore, employing modern methods such as classical and neural network-based designs, the construction of creative ransomware cures can be accomplished effectively. On a chosen set of attributes for categorizing ransomware, investigators applied a range of machine learning techniques, including Random Forest (RF), Logistic Regression (LR), SVM, KNN, and decision tree. To evaluate the suggested strategy, we ran each test on a single ransomware sample. For example, ransomware frequently rushes through a variety of document-related activities in order to lock or encrypted the files on a victim's computer. Users' data can't be effectively protected against assaults carried on by hazardous unrecognized ransomware when using signature-based malware detection techniques, due to issues identifying zero-day ransomware.

Keywords— *RF-Random Forest, LR-Logistic Regression, API- Application Programming Interface, C&C- Commander and Control, SVM-Support Vector Machine.*

I. INTRODUCTION

In order for increasingly high-level and complex ransomwares to spread throughout the world and have a substantial impact on people, businesses, governments, and entire nations, ransomwares are continuously developed in hidden markets. In order to prevent widespread attacks, most companies spend money on detection systems for intrusions that alert them of any strange network behavior. Despite being related to ransomwares, they are not detectable using traditional detection based on signatures. The platform for digital cash known as bitcoin is used to pay ransomware. The first ransomware had a symmetric-key method, which required the same key to decode and encrypt data. The key is provided to the C&C server after encrypting is complete. The decryption key may be taken as it has been transmitted across several networks [1]. The majority of current ransomware accepts Bitcoin as compensation. Although Bitcoin transactions are publicly accessible and continuously recorded, current methods for identifying ransomware solely rely on a few arduous gathering of data processes and/or heuristics (such as executing ransomware to acquire Bitcoin addresses associated to ransomware). To our knowledge, no prior methods have used cutting-edge statistical methodologies to instinctively recognize payments connected to malware and malicious Bitcoin accounts [2]. Bitcoin transactions can be performed privately, and verification of identity is not necessary to join the network. A payment can be requested by sending an individual a short string representing a public Bitcoin address across anonymity networks like Tor. Malicious actors have taken notice of Bitcoin's simplicity of use and global transaction accessibility.

The ransomware detection method proposed in the present research may distinguish between ransomware and malware as well as among harmful software and safe data. Using the Intel PIN tool, they collected Windows API (Application Programming Interface) call patterns, and from these sequences, n-gram sets were produced [3]. In order to lock files for victims, ransomware typically performs a lot of file-related actions in a short amount of time. Additionally, it might be challenging to decrypt information without decryption keys after a machine has been attacked with ransomware and some files have been encrypted [4]. Two types of ransomware may be distinguished: The ransomware Locker Next, we have crypto-ransomware. The primary mechanism of ransomware is identical in both groups, but the lockable ransomware prevents the user from using the machine, which is the difference between the two types. On the other side, the crypto ransomware will use cryptographic operations to scramble the system's data and directories [5].

Recent ransomware outbreaks have shown that it may be a way for attackers to make money. Because antivirus software is unable to recognize sophisticated advanced persistent threat attacks or unknown malware, it is utilized to thwart ransomware attempts. Malware has the capacity to blind, evade, tamper, and alter its behavior in order to appear innocent when being inspected, especially sophisticated malware that runs on the same system OS. Additionally, it can turn off the entire machine to prevent detection. [6]. Computing system security is significantly threatened by ransomware. As a result, detection of ransomware has gained popularity in the field of technological security. Deception, variation, enlargement, and encrypting can, nevertheless, regularly be used to readily evade the present signature-based and static detection approaches [7].

Ransomware, which demands payment in exchange for access to one's files, continues to grow despite the introduction of numerous surveillance and avoidance techniques to safeguard user data. Utilizing file- and behavior-based detection techniques, ransomware has been investigated for recognition and prevention. However, ransomware assaults continue to occur, partly because it is difficult to identify and halt malware that incorporates unidentified dangerous software. The inability of these techniques to identify ransomware for backup services in the cloud or other backup options is one of its many drawbacks [8]. The Bitcoin crypto currency was created in 2008 as a distributed transactions system and is now a widely used virtualized digital currency. Without a middleman, peer-to-peer network nodes are used in Bitcoin transactions, and the node can verify the transactions. Although the effectiveness of the Bitcoin networks in terms of financial transaction systems has been high, their financial transactions are susceptible to a number of ransomware assaults. In order to prevent such damaging intrusions, researchers have been striving to build ransomware payment recognition tools for bitcoin transaction systems [9].

A type of malware known as ransomware is increasingly posing a severe online danger to both individuals and businesses worldwide. Ransomware criminals, in contrast to regular malware, take over the machine and demand money to undo the attack. It is more challenging to accurately identify and categories these attacks because attackers can use polymorphic, metamorphic, and other masking techniques to create new strains of malware that is already in existence. Traditional static analysis methods are losing their ability to identify emerging ransomware variations, categories them, and offer insight into the danger, objectives, and behaviors of ransomware. Technologies for behavior-based classification have been developed. There is a growing need for greater research in categorization methods because to the rapid proliferation and diversification of ransomware in recent years [10]. Attacks using ransomware are on the rise at the moment. Many governmental and non-governmental organizations, particularly those in the fields of education, wellness, finance, science, and assurance, have been impacted. The user's computer and data have been taken over using methods including social engineering attacks, cracking passwords, hacking of networks, and others in order to inflict more harm and disruption. The unavailability of the live system, disruption of routine operations, cost of forensic analysis, cost of reconstruction, losses due to reputational

damages, and the price of cyber security instruction, ransomware assaults result in substantial costs of damages [11].

Cybercriminals have taken notice of this older piece of software due to its successful attack and rapid financial gain. By encrypting the computer's operating system or encrypting particular files that the victim considers critical, such as images, spreadsheets, and slideshows, ransomware aims to stop the victim from accessing their own resources. The two primary ransomware subcategories are locky and crypto. [12]. Even though Locky ransomware blocks entry to every part of the machine, it is frequently simple to remove. Contrarily, crypto ransomware locks off specific files from user access via cryptography; this is much more difficult to fix, and the harm could be catastrophic. Cybercriminals most frequently employ crypto ransomware as a form of ransomware. Scareware is a different type of ransomware that is referenced in the source material. The victim's machine is not truly affected by the ransomware; they are just coerced into paying the ransomware [13].

The ransomware seeks to lock the user's computer using simple or sophisticated procedures, making it impossible for the user to regain entry to the device. Then, they regularly flash an email requesting payment across the screen. Access is only reestablished after the ransom has been paid. On the other hand, crypto-ransomware hunts down user files, discreetly encrypts them, and then demands a ransom in exchange for the codes that are required to regain accessibility to those files [14]. Ransomware is a specific type of virus that causes irreparable data loss and has high financial consequences. Nowadays, detecting ransomware is
a crucial effort. Some ransomware may track the run-time environment and avoid dynamic analysis, such as fingerprinting malware. In order to identify this sort of malware and deal with data more quickly than with dynamic analytics [15].

II. LITERATURE SURVEY

Dragos *et al* (2019) [1] have argued that the primary goal should be to develop a machine learning framework that can broadly identify as many variants of malware as possible while adhering to the strict requirement of zero false positives. This structure needs a variety of predictable exceptions techniques to be included in an extremely competing business offering. The cascade one-sided the perceptron (COS-P) and its explicitly mapped counterpart (COS-PMap) constitute the most dependable techniques among those provided here because every commercial anti-virus programme is bound to certain performance and storage restrictions.

Mohammad Masum *et al* (2022) [2] have incorporated various machine learning techniques, particularly neural network-based classification algorithms, and proposed an element selection-based innovative system for efficient ransomware classification and detection. On a ransomware data set, we used the framework along with all of the trials, and we assessed the efficacy of the models using a thorough comparison of DT, RF, NB, LR, and NN classifications.

Umme Zahoora et al (2022) [3] although the architecture that has been presented takes host-based features into account, network traffic authentication may also be

studied as a feature in the future. The longer version of the present study, which focuses on eleven kinds of ransomware, might involve retraining the suggested framework using other blackmail variations. Dynamic evaluation takes time to complete. By focusing primarily on pre-encryption based characteristics in the future, this characteristic's extraction time can be decreased.

The data set processes a large amount of data before sending the output to the preprocessing unit. It splits the data after processing it. There are two categories within the data split. Test Data and a trained set. Compared to test data, the trained set has a large amount of data. Data from the trained set is transferred to the model, which compares the outcome to the trained set. Attack and non-attack data are available for Trained Set. When this data are compared with attack data, the result is an attack. Otherwise, it produces no effect.

III. PROPOSED SYSTEM DESCRIPTION

Every time the process occurs, the ransomware detection algorithm based on deep learning extracts the API sequences to process and generates n-gram sequences. Tenfold cross-validation was utilized using input files from deep learning algorithms that were split into training and testing sets. The classification system will be created in Python as a Jupiter kernel for stream processing.

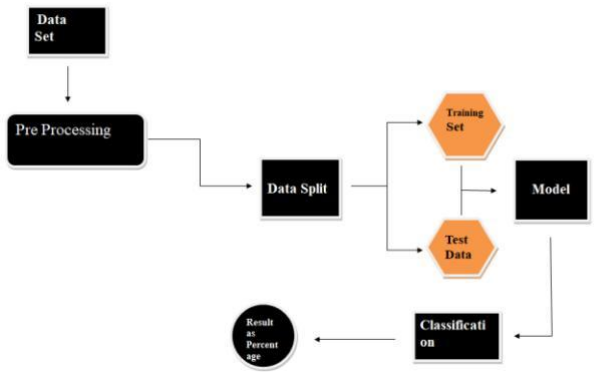

Fig. 1. Proposed Diagram

The amount of data's are flow through the data set it transfers the output to the preprocessing unit. It process the data to data split. Trained set has high amount of data's and test data acquires low amount of data. Trained set data transfer the output to model it compares the result to trained set. Trained Set have attack and non-attack data's. By compare these data result if it compares with attack data it has attack output. Otherwise it has non attack output.

IV. PROPOSED SYSTEM MODELLING

A) RANDOM FOREST

Among the best artificial intelligence methods is random forest. There should be duplicates for about 37% of the occurrences in the bootstrap data sets that were created. Another kind of randomization utilized in random forest includes attribute selection. To find the best split, a subset of the input parameters is chosen at random for each node split. Breiman suggests that this variable be given the value "log2 (#features) +1." For categorization, majority voting determines the ensemble's final forecast.

The more trees there are in the forest, the more likely it is that the ensembles has crossed the asymptotic generalization error. The fact that random forest is practically parameter-free, or at least performs admirably on average with its standard parameter value, is also one of its key advantages. The most effective two approaches in that comparative analysis are based on random forests, where tuning is limited to the quantity of random attributes chosen at each split. Random forest with the standard setting came in sixth (out of 179 methods) in the evaluation. This may also be viewed as a disadvantage because parameter adjustment for random forest is challenging.

In the present investigation, we assess a number of variables from the group that can be modified for random forest:

The maximum amount of features to take into account while determining the optimum split.

• The tree's deepest point (max_depth). No matter how many examples are contained in each node, this value restricts the length of the tree.

B) LOGISTIC REGRESSION

To investigate the effect of predicting components on subcategory answers, logistic regression models are employed. A model of logistic regression is known as simple logistical regression when there is just one predictor variable. The multivariate or multi-variable logistic regression method, which includes both categorical and continuous variables as predictors, is used when there are many predictors (such as risk factors and treatments).

Logistic algorithms are commonly employed in epidemiological research to examine the relationships between risk variables and disease development. Such frameworks are frequently used for medical articles that do not focus on epidemiology and public health. The logistically method is the sophisticated statistical framework (models that correct for confounders) that is commonly used in medical journals of significant significance to their area of research.

C) SUPPORT VECTOR MACHINE

SVM, is one of the most widely used methods for supervised data mining for resolving challenges associated with classification and regression. Nevertheless, it is usually utilised to solve challenges with Machine Learning Categorization. The SVM method aims to locate the most effective line or decision borders that may split the space of n dimensions into categories in order to quickly categorise brand-new information in the near future. The highest judgements border is designated as a hyperplane. SVM selects the points and maximum vectors that contribute to the formation of the hyperplane. The SVM technique is based on support vectors, which are used for modelling these key scenarios.

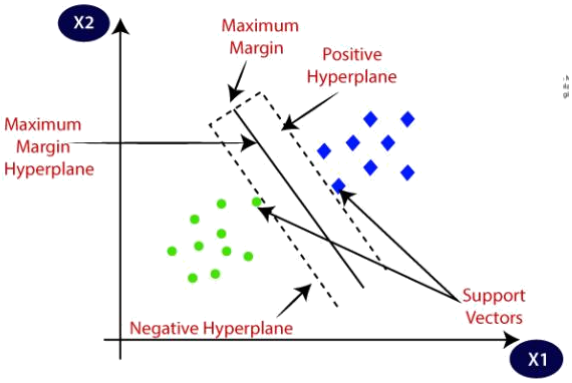

Fig. 2. Support Vector Machine

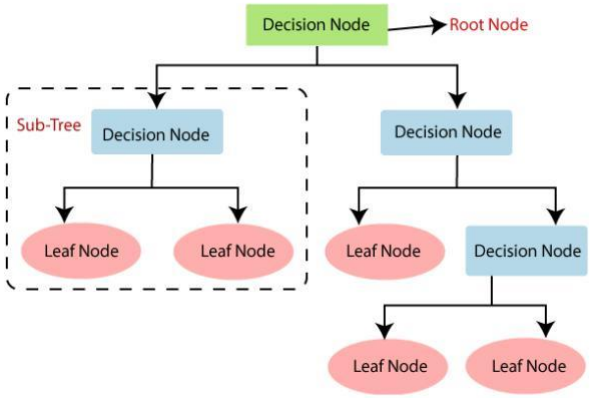

Fig. 3. Decision Tree Algorithm

SVM can be of two types:

Linear SVM: Linear SVM is used to separate data into two separate categories using a single continuous line. This kind of information is referred to as data that is linearly separate, and the technique used is known as a Linear SVM predictor. Non-linear SVM: Non-linear SVM is used for data that does not divide uniformly.

D) KNN

KNN is a straightforward and reliable categorization method. The training and testing data vectors are separated by Euclidean, cityblock, cosine, and correlation distances in this approach.

(1)

The testing vector receives the label from the characteristic vector with the least separation. Then, using a proper metric, we can relate the gap between two distinct points in a space to how similar they are to one another. The following is the K-nearest neighbor technique. 1. The new sample and an optimistic integer value k are defined. 2. Choose our database's k values that are most similar to the new testing sample. 3. We determine which category of these comments is the most comparable. 4. Using the value of k, we categorise the new sample as follows. 5. The value of k was modified until the desired results were not achieved.

E) DECISION TREE

Although a decision tree, a supervised learning method, can be used to solve issues with regression or classification, it is often preferable. It is a tree-structured classifier, with internal nodes reflecting data set properties, branches representing the decision-making process, and every node in the leaf indicating the classification outcome.

Step 1: Begin the tree at the root node, which contains the complete dataset, in accordance with S. Step 2: Use the Attribute Selection Measure to determine the top attribute in the dataset. Step 3: Divide the S into sections to provide various possibilities for the finest attributes. In step four, create the decision tree node with the best attribute. In step 5, continuously create new decision trees using the data set choices produced in step 3.

IV. RESULTS AND DISCUSSION

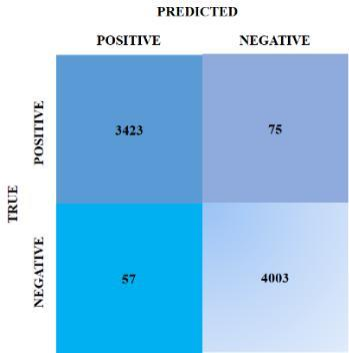

Fig. 4. Decision Tree Confusion Matrix

Choice Tree X train values are provided as input for the decision tree classification during training. Once the learning procedure is finished, the decision tree technique's correctness is examined using X test values. The decision tree classification achieves 99.14% accuracy. The matrix of confusion for the Decision Tree Classifier is displayed in picture 4.

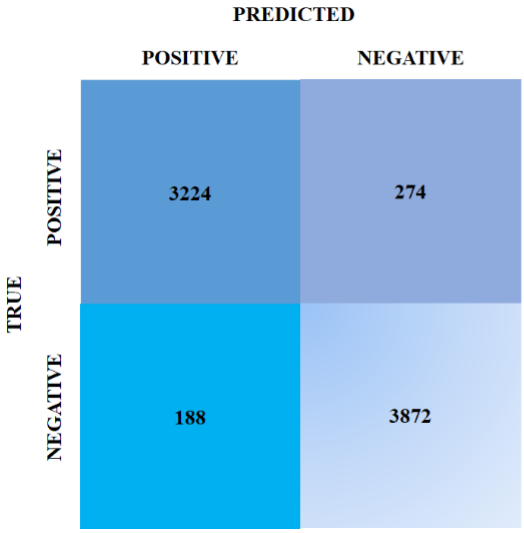

Fig. 5. Random Forest Confusion Matrix

Rough Forest X train values are provided as input for the random forest classifier during learning. While training with Random Forest Classifier takes longer than with Decision Tree and Light GBM, the accuracy gained is higher than with Decision Tree but less than with Light GBM. After the learning process is finished, the Random Forest technique's correctness is examined using X test values. The precision of the Random Forest classifier is 99.47%. The matrix of confusion for the Random Forest Classifier is displayed in Figure 5.

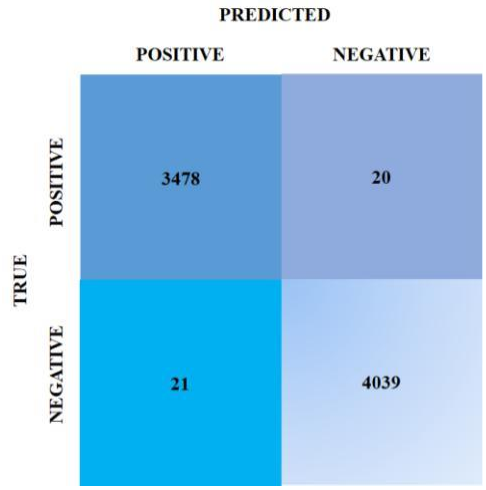

Fig. 6. Light GBM Confusion Matrix

For instruction, Light GBM Input X train values are offered. Light GBM trains more quickly than Random Forest and Decision Tree methods. Of the other two, the accuracy attained is also the highest. Utilising X test values, the Light GBM algorithm's accuracy is evaluated following its training procedure. 99.50% of the time, the Light GBM categorization is accurate. Figure 6 shows the muddled matrix for the Light GBM Divider.

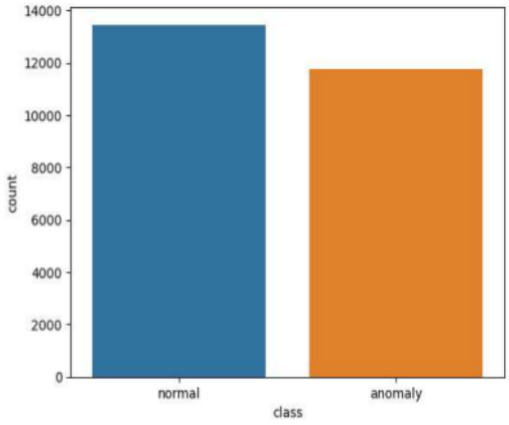

Fig. 7. Class

Figure 7 compares the count between normal and anomaly. End of the comparison the normal count is higher.

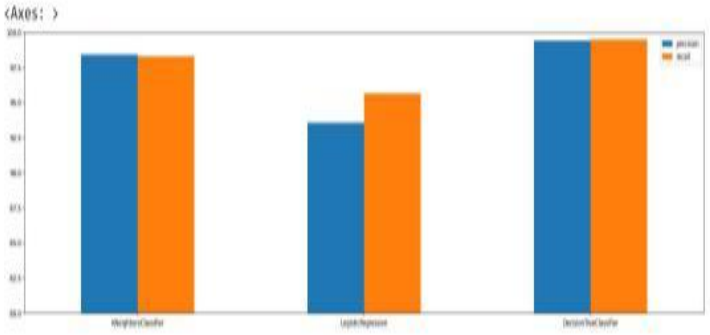

Fig. 8. Classification

In comparison of KNeighbors Classifier, logistic regression and decision tree classifier received by decision tree classifier is better and the comparative results will displayed in the figure 8. Hence, decision tree classifier has the best classification.

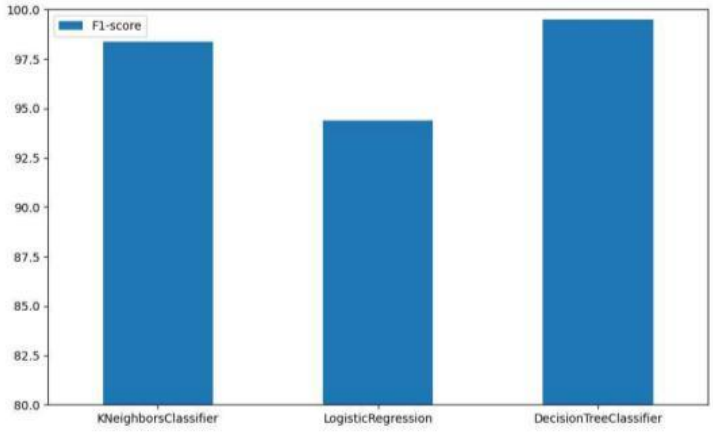

Fig. 9. Comparison Graph

In comparison of KNeighbors Classifier, logistic regression and decision tree classifier, the decision tree classifier. The decision tree classifier has the greater value.

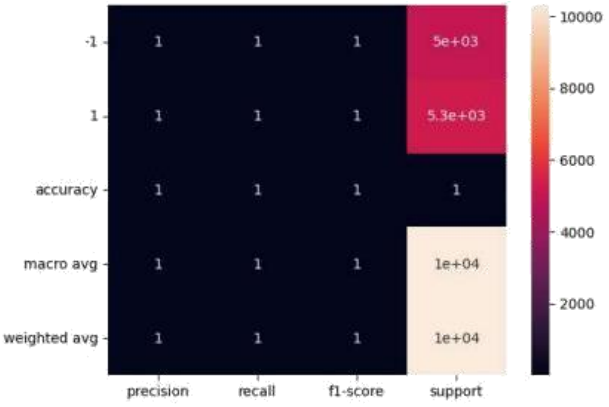

Fig. 10. Decision Tree

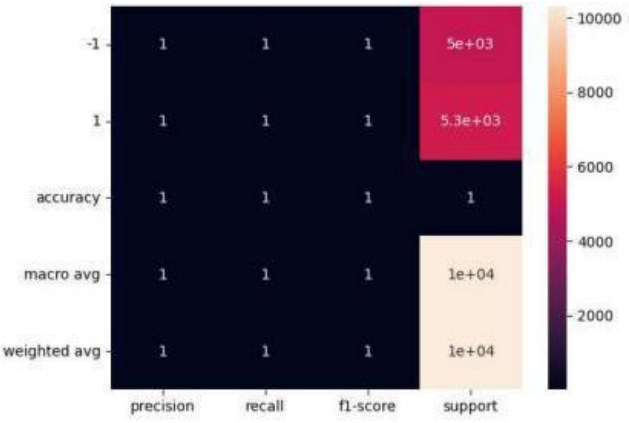

Fig. 11. Logistic Regression

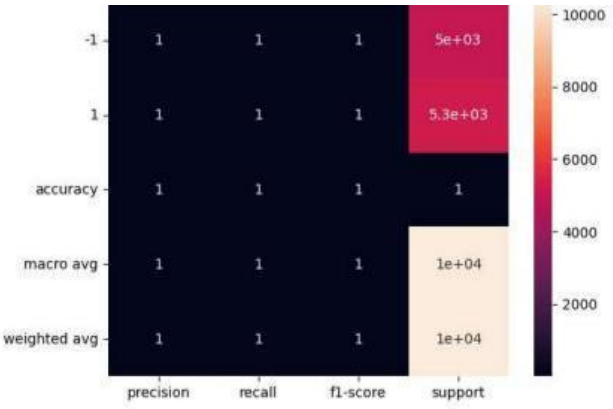

Fig. 12. SVM

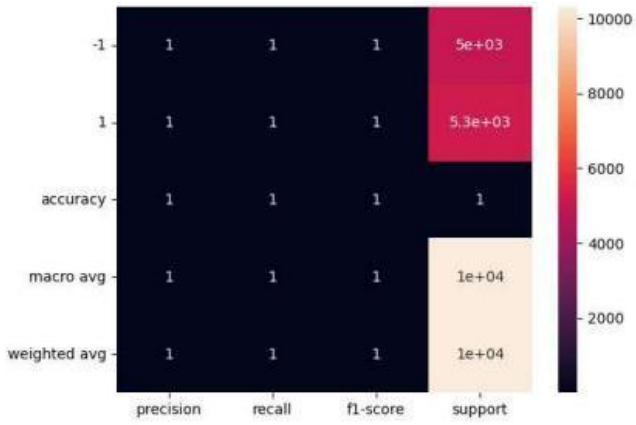

Fig. 13. Random Forest

The Confusion matrix of the Decision Tree, Logistic Regression, SVM and Random forest in which they are listed in Fig. 10,11,12 and 13 Respectively.

V. CONCLUSION

The capacity for ransomware groups to substantially impair computer systems, data centers, websites, and mobile apps in a number of businesses and professions makes them serious security risks for cybersecurity. Conventional anti-ransomware software developers struggle to counter newly created, complex threats. Consequently, the development of innovative ransomware solutions can be successfully achieved using contemporary techniques like as classical and neural network-based designs. We used a variety of machine learning methods, such as decision tree, Logistic Regression, SVM, KNN, and Random forest on a selected set of characteristics for categorizing ransomware.

REFERENCES

1. Wan, Yu-Lun, Jen-Chun Chang, Rong-Jaye Chen, and Shiuh-Jeng Wang. "Feature-selection-based ransomware detection with machine learning of data analysis." In 2018 3rd international conference on computer and communication systems (ICCCS), pp. 85-88. IEEE, 2018.
2. Akcora, Cuneyt Gurcan, Yitao Li, Yulia R. Gel, and Murat Kantarcioglu. "Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain." arXiv preprint arXiv, pp. 1906.07852, 2019.
3. Tsoutsanis, Panagiotis. "Extended bounds limiter for high-order finite-volume schemes on unstructured meshes." Journal of Computational Physics, Vol. 362, pp. 69-94, 2018.
4. Bae, Seong IL, Gyu Bin Lee, and Eul Gyu Im. "Ransomware detection using machine learning algorithms." Concurrency and Computation: Practice and Experience, Vol. 32, no. 18, pp. e5422, 2020.
5. Vinayakumar, R., K. P. Soman, KK Senthil Velan, and Shaunak Ganorkar. "Evaluating shallow and deep networks for ransomware detection and classification." In 2017 international conference on advances in computing, communications and informatics (ICACCI), pp. 259-265. IEEE, 2017.
6. Adamu, Umaru, and Irfan Awan. "Ransomware prediction using supervised learning algorithms." In 2019 7th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 57-63. IEEE, 2019.
7. Chen, Zhi-Guo, Ho-Seok Kang, Shang-Nan Yin, and Sung-Ryul Kim. "Automatic ransomware detection and analysis based on dynamic API calls flow graph." In Proceedings of the international conference on research in adaptive and convergent systems, pp. 196-201. 2017.
8. Lee, Kyungroul, Sun-Young Lee, and Kangbin Yim. "Machine learning based file entropy analysis for ransomware detection in backup systems." IEEE Access, Vol. 7, pp. 110205-110215, 2019.
9. Al-Haija, Qasem Abu, and Abdulaziz A. Alsulami. "High performance classification model to identify ransomware payments for

heterogeneous bitcoin networks." Electronics, Vol. 10, no. 17, pp. 2113, 2021.

10. Daku, Hajredin, Pavol Zavarsky, and Yasir Malik. "Behavioral-based classification and identification of ransomware variants using machine learning." In 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE), pp. 1560-1564. IEEE, 2018.
11. Poudyal, Subash, Dipankar Dasgupta, Zahid Akhtar, and Kishor Gupta. "A multi-level ransomware detection framework using natural language processing and machine learning." In 14th International Conference on Malicious and Unwanted Software" MALCON, no. October 2015. 2019.
12. Fernandez Maimo, Lorenzo, Alb erto Huertas Celdran, Angel L. Perales Gomez, Felix J. Garcia Clemente, James Weimer, and Insup Lee. "Intelligent and dynamic ransomware spread detection and mitigation in integrated clinical environments." Sensors, Vol. 19, no. 5, pp.1114, 2019.
13. Kok, S., Azween Abdullah, N. Jhanjhi, and Mahadevan Supramaniam. "Ransomware, threat and detection techniques: A review." Int. J. Comput. Sci. Netw. Secur, Vol. 19, no. 2, pp. 136, 2019.
14. Sgandurra, Daniele, Luis Muñoz-González, Rabih Mohsen, and Emil C. Lupu. "Automated dynamic analysis of ransomware: Benefits, limitations and use for detection." arXiv preprint arXiv, pp. 1609.03020, 2016.
15. Zhang, Bin, Wentao Xiao, Xi Xiao, Arun Kumar Sangaiah, Weizhe Zhang, and Jiajia Zhang. "Ransomware classification using patch-based CNN and self-attention network on embedded N-grams of opcodes." Future Generation Computer Systems, Vol. 110, pp. 708-720, 2020.