Detection Of Ransomware Using Machine Learning Algorithm Based on PE Header Features

Sudeep K S[1], Sri Ram B K[2], Vinod Mallappa Hugar[3], Yashwanth S[4], Mr. Shravan H S[5]
[1,2,3,4,5]*Dept.of.CS&E, PESITM Shimoga*

***Abstract*—This research is about the detection of the ransomware files using support vector machine (SVM) which uses pe header features as a dataset to classify the files as ransomware or a benign file.The model can classify a pe file based on static analysis also achieves a great accuracy. An additional feature combining regex-based extraction of ASCII and Unicode strings from the PE files with keyword-based matching to detect potential ransomware notes.**

***Index Terms*—Ransomware, Benign, Support Vector Machine (SVM), Machine learning, Cyber security, PE file, ASC II, Unicode, ransomware notes.**

I. INTRODUCTION

Ransomware is a type of software which is designed to encrypts the data from the victim's computer and demands for ransom. The software program that executed on a victim's computer encrypts all the data of the victim. The decryption key is provided to the victim only if the ransom is paid. Otherwise, all the data of the victim may lose permanently which may contains all the financial details of the victim and also personal data. So, this ransomware became a challenge for cybersecurity since the start of the 1980.

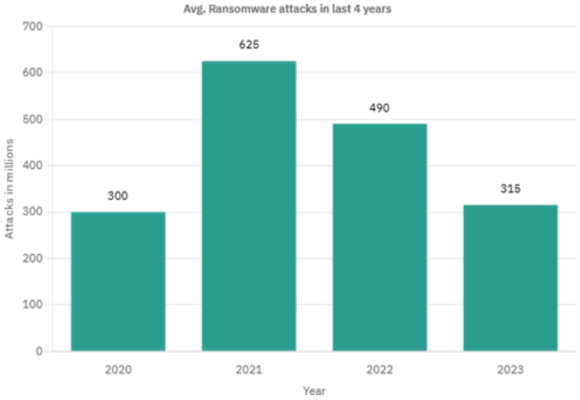


Fig 1: Ransomware attacks in last 4 years

The above graph shows the average ransomware attacks every year which indicates the severity of ransomware attacks from the last 4 years excluding year 2024.

Let alone in 2021 there were around more than 600 million attacks globally which is a major threat to the cybersecurity. It is also being said that there is an increase in 10% of data breach from ransomware attacks from reports of IBM. JumpCloud an IT platform stated that ransomware attacks are increasing intensely as compared to last few years.

In order to overcome this threat so many researchers as well as scientists found many ideas to prevent this infection. Some algorithms and methods have achieved great accuracy by using machine learning algorithms, natural language processing, static analysis, anomaly-based analysis of the files, network-based analysis and also based on behavioral analysis. Most of the study as well as research is mainly based on behavioral analysis and network analysis. So, this paper mainly focuses on detecting the ransomware file based on the features of PE header of the PE (portable executable) file.

So why executable files? why not the other methods?

The main reason is that the features of the pe file can be extracted without actually pe file being executed. The detection time will be much higher when compared to other methods. The resources that are required by using pe header features within pe file is much lower and also the risk of infection of the virus to the system is also minimal. Other than analysis of ransomware using natural language processing of ransom notes, pe header analysis method has minimal false positive or false negative.

II. RELATED WORKS

Many research and study have been carried out and is being studied even today about the behavior of the ransomware and how it can be prevented without being infected by that virus. Some literature survey carried out by different researchers are as follows and how it affects the current cybercrimes and also which topic can be further studied to avoid those mistakes and also to update them.

Research carried out by [1] R K Bansal and S V Rao in 2018 about detecting ransomware using hybrid machine learning models that combines clustering and classification algorithms such as changes in file system, pattern of data that are being encrypted and also how the process of that particular execution by using both supervised and unsupervised learning. This combination of using both supervised and unsupervised learning may lead to the requirement of large dataset as compared to static analysis and also the algorithm is very complex to generate and may increase detection time. Although the accuracy may hold high but implementation is quite challenging.

The study carried out by [2] Krishna Kant and Kiran S in the year 2020 about ransomware detection using machine learning techniques to monitor behavior. This study shows that we can use multiple algorithms like SVM and Random Forest to monitor behaviors like resource utilization and file operation. So, the problem here is that the ransomware file may encrypt data even before monitoring these behavioral changes. So, what is the use of applying multiple models if the injection of the virus is already done?

Here comes the answer to that question, that we can use static analysis instead of monitoring the behavioral changes. The model may be effective but it has a lot of risk and also ransomware is not like other malwares. The decryption key of the data is with the threat actor and if it encrypts your file you have to pay ransom in order to revive those data. In this case also static analysis of pe header features of the pe file comes into picture.

Ransomware detection using behavioral based machine learning technique, a study carried out by

[3] Yassir M Alharthy shows that we can classify the file whether it is a normal file or a ransomware file using KNN clustering and Naïve Bayes (NB) algorithm based on the system calls and file operations. Firstly, KNN requires comparing of test data to all the training instances in the dataset to find the nearest neighbor of that test data this might take time for more features and also for high dimensionality data this may increase the detection time. In Naïve Bayes it considers every feature to be non-dependent on other features which may be false and this can lead to the false negative or false positive.

The research carried out by [4]M. A. Moustafa and K. A. Moustafa highlights mainly on behavioral analysis categorizing them based on signature and behavior of the ransomware which may increase the possibility of risk for victim's computer as it based on behavior of ransomware which in turn often requires execution of that file and also Comprehensive review of ransomware detection and mitigation using machine learning by[5] Rajeev M S ,Pradeep K S and Shashank G in 2020 shows that the authors had reviewed different strategies ,static ,dynamic and also hybrid approaches to address evolving characters and also the review also addresses the various machine learning algorithm like Random Forest, Support Vector Machine(SVM) and Decision tree in ransomware detection based on behavioral file analysis and network traffic .In this there might be a lack of dataset to satisfy all the algorithm and there is an equal proportion of probability that the virus may get injected even before it detects the behaviors .So it is still a risky move as compared to static analysis of the file that extracts the features without actually executing the file. This factor also plays a major role in selecting static analysis compares to behavioral or any other analysis.

III. METHODOLOGY

A. Collection of the Data

The dataset used in this research contains both benign as well as ransomware pe file features like DOS header, debugrva, size of stack reserve, optional headers, section headers etc, these are the features of the PE header and can be extracted from every normal pe file. This dataset contains around 60000 features of PE header of both ransomware as well as benign in equal proportion.

B. Features Extraction

These features of the PE file now extracted using python library PE file which is an inbuild library for extracting pe file features or details. Now converting categorial data to numerical data using different encoding like one hot encoder and also removing columns of the dataset that are not required by the model and finally removing all the null values and followed by extraction of the features, data preprocessing takes place.

C. Data Preprocessing

In Data Preprocessing the dataset in which the categorial data is converted into numerical data is now have to normalize or scalable which is an important step in this methodology due to which the performance of the SVM may alter. So, it is necessary to preprocess the data before the testing of the data. Now splitting the data set into 80% training and 20% for the testing phase which trains the model to detect the testing data based on the features of the trained dataset.

D. Feature Selection

The main objective of the feature selection is to select only the relevant features from the pe header which helps the model to increase the efficiency by removing the irrelevant data. Features that are closely related to one another and the features that are irrelevant and insignificant can be dropped from the data. The model accuracy and efficiency mainly dependent on the features that are significantly contribute to the dataset. So, removing other than such features increases model efficiency and accuracy.

E. Model Training

Training the model using Support Vector Machine (SVM) algorithm because of its effectiveness for even minimal number of the features. Now a SVM kernal is used to separate the dataset into two categories as both benign and ransomware files based on the features of the pe header provided in the dataset. Using the training dataset to classify the data into ransomware or a benign file. As our dataset contains equal proportion of both ransomware as well as benign files handling class imbalance has less priority. You can use Class Weighting as well as SMOTE for handling class imbalance. SVM identifies specific patterns of the ransomware files that are differ from the benign files and vice versa. Training SVM model to its highest potential increases the overall performance of the model.

F- Evaluation of the Model.

Evaluation of the model is the final and important step where the data from the outside world is tested using SVM model which is trained by the same features in a given dataset. This step contains evaluating accuracy, precision etc., to check how effective model works in the real time. Evaluating true positives, true negatives, false positives and false negative to evaluate the performance of the model against a test data in real time which shows the model's real-world performance. The main objective of evaluating the model is to cross verify its robustness, accuracy, detection of ransomware files in real time, to alter the features for upcoming ransomware family pe features etc,.

F. Detection of Ransom Note

An additional feature of detection of ransomware note using the methods combining regex based extraction of ASCII and Unicode strings from the PE file with keyword-based matching to check for suspicious ransom notes or words like decrype,ransom,bitcoin,encrypt etc.,

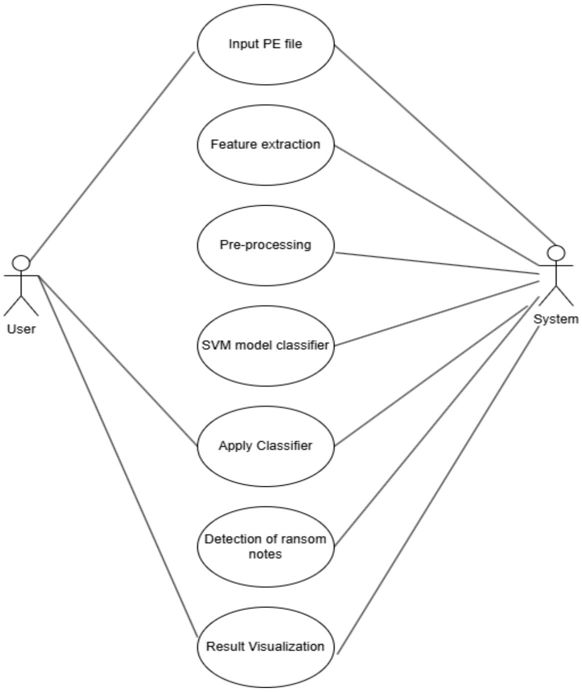


Fig 2: Use Case diagram.

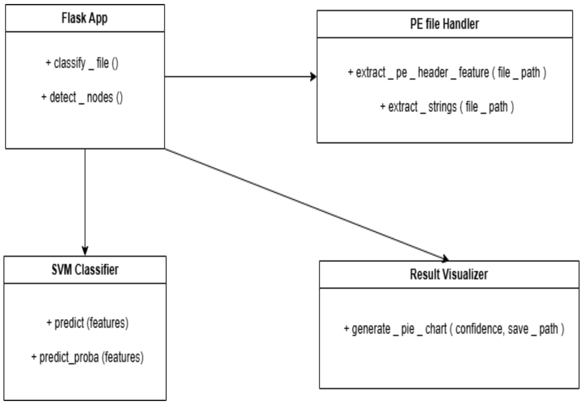

Flask App
+ classify _ file ()
+ detect _ nodes ()
PE file Handler
+ extract _ pe _ header _ feature (file _ path)
+ extract _ strings (file _ path)
SVM Classifier
+ predict (features)
+ predict_proba (features)
Result Visualizer
+ generate _ pie _ chart (confidence, save _ path)

Fig 3: Class diagram

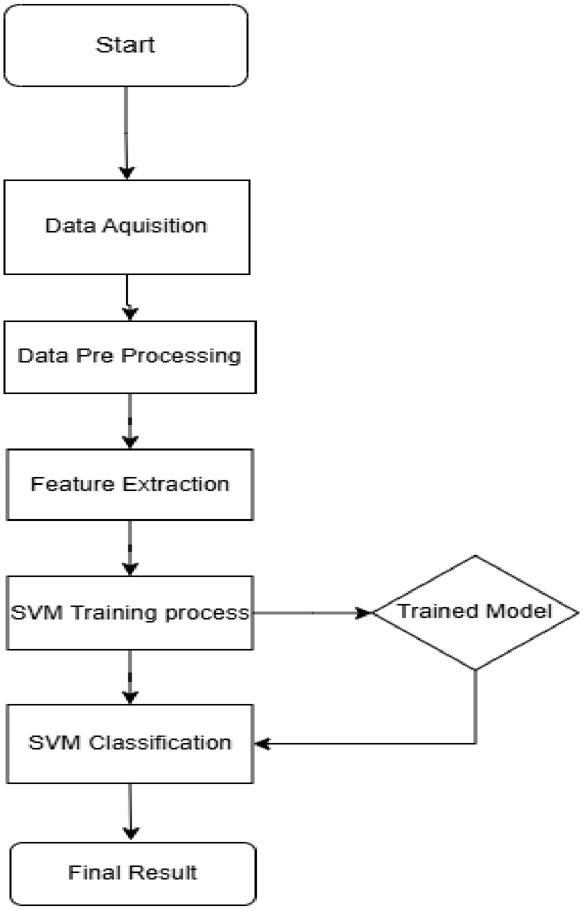


Fig 4: Flow Chart for SVM model

IV. RESULT ANALYSIS

Support Vector Machine (SVM), a machine learning algorithm is used. PE file is added and features are extracted from a PE file for preprocessing and using SVM model the file is classified as ransomware or benign. A ransom note detection, an additional feature is also added to detect ransom notes using regex-based methodology by extracting ASCII and Unicode strings for suspicious words.

The model achieves promising results in prediction of ransomware notes using SVM model.

- Confidence Score: Indicates the confidence level of the classification.
- Model Accuracy: Represents the overall accuracy of the underlying model.
- Classification Report: Shows precision, recall, F1-score and support metrices.
- Risk Factor: A pie chart showing risk of opening that file.

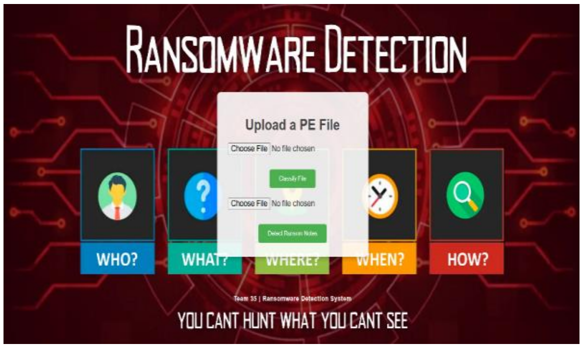
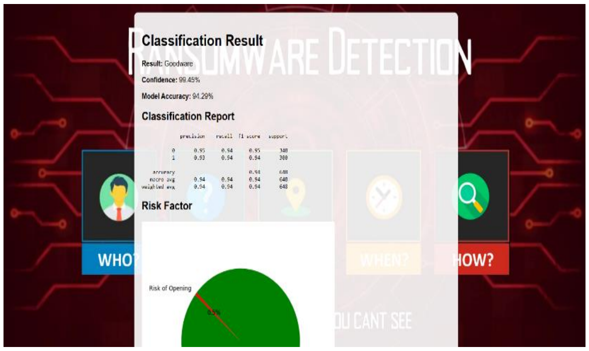

RANSOMWARE DETECTION
Upload a PE File
Choose File No file chosen
Classify File
Choose File No file chosen
Detect Ransom Notes
WHO?
WHAT?
WHERE?
WHEN?
HOW?
Team 35 | Ransomware Detection System
YOU CANT HUNT WHAT YOU CANT SEE

Fig 5: User Interface for uploading pe files.



Classification Result

Result: Goodware

Confidence: 99.45%

Model Accuracy: 94.29%

Classification Report

```
              precision    recall  f1-score   support

           0       0.95      0.94      0.95       340
           1       0.93      0.94      0.94       300

    accuracy                           0.94       640
   macro avg       0.94      0.94      0.94       640
weighted avg       0.94      0.94      0.94       640
```

Risk Factor

Fig 6: Classification results.

V. CONCLUSION

In this study we conclude that we achieved highest accuracy ranging from 94.29% to 99 % for all types of features of the PE header of a PE file. Static Analysis features low infection risks and high efficiency and accuracy and is effective for the PE files without being executed. Support Vector Machine (SVM) is effective for this static analysis and also for additional ransom note detection a regex-based extraction of ASCII and Unicode strings for detection of suspicious ransomware related words or notes. Thus, the efficiency and also the accuracy of the model is high and has low risk factor. Our study concludes that in a faster detection time and low risk factor we can classify PE files as ransomware or benign along with detection of ransom notes as additional feature.

REFERENCES

[1] R. K. Bansal and S. V. Rao, "Detecting ransomware using hybrid machine learning models," in Proceedings of the Conference on Emerging Security Technologies, 2018.

[2] K. Kant, K. S. Bhat, et al., "Ransomware detection using machine learning techniques," in

International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 3, pp. 123-130, 2020.

[3] Y. M. Alharthy, et al., "Ransomware detection using behavior -based machine learning techniques," in Journal of Cybersecurity Research, vol. 8, no. 4, pp. 45-56, 2021.

[4] M. A. Moustafa and K. A. Moustafa, "A comprehensive survey on ransomware detection techniques: From signature-based to machine learning based approaches," in IEEE Access, vol. 8, pp. 123456-123470, 2020.

[5] R. M. S., P. K. S., and S. G., "A comprehensive review of ransomware detection and mitigation using machine learning," in ACM Computing Surveys, vol. 52, no. 3, pp. 1-20, 2020.